Express Mail No.: EL769974062US
Attorney Docket No.: 018501-001200US
EOS Docket No. BRCA-001-3

PATENT APPLICATION

METHODS OF DIAGNOSIS OF BREAST CANCER, COMPOSITIONS AND METHODS OF SCREENING FOR MODULATORS OF BREAST CANCER

Inventor(s):

David H. Mack, a citizen of the United States, residing at 2076 Monterey Avenue, Menlo Park, California 94025

Kurt C. Gish, a citizen of the United States, residing at 40 Perego Terrace #2, San Francisco, CA 94131

Assignee:

EOS Biotechnology, Inc.
225A Gateway Boulevard
South San Francisco, California 94080-7019

Entity: Small

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 415-576-0200

PATENT
Express Mail No.: EL769974062US
Attorney Docket No.: 018501-001200US
EOS Docket No. BRCA-001-3

METHODS OF DIAGNOSIS OF BREAST CANCER, COMPOSITIONS AND METHODS OF SCREENING FOR MODULATORS OF BREAST CANCER

CROSS-REFERENCES TO RELATED APPLICATIONS

This application is a continuation-in-part of USSN 09/525,361, filed March 15, 2000, which is incorporated herein by reference.

FIELD OF THE INVENTION

The invention relates to the identification of nucleic acid and protein expression profiles and nucleic acids, products, and antibodies thereto that are involved in breast cancer; and to the use of such expression profiles and compositions in the diagnosis, prognosis and therapy of breast cancer. The invention further relates to methods for identifying and using agents and/or targets that inhibit breast cancer.

BACKGROUND OF THE INVENTION

Breast cancer is one of the most frequently diagnosed cancers and the second leading cause of female cancer death in North America and northern Europe, with lung cancer being the leading cause. Lifetime incidence of the disease in the United States is one-in-eight, with a 1-in-29 lifetime risk of dying from breast cancer. Early detection of breast cancer, using mammography, clinical breast examination, and self breast examination, has dramatically improved the treatment of the disease, although sensitivity is still major concern, as mammographic sensitivity has been estimated at only 60%–90%. Treatment of breast cancer consists largely of surgical lumpectomy or mastectomy, radiation therapy, anti-hormone therapy, and/or chemotherapy. Although many breast cancer patients are effectively treated, the current therapies can all induce serious side effects which diminish quality of life. Deciding on a particular course of treatment is typically based on a variety of prognostic

parameters and markers (Fitzgibbons et al., 2000, Arch. Pathol. Lab. Med. 124:966-978; Hamilton and Piccart, 2000, Ann. Oncol. 11:647-663), including genetic predispostion markers BRCA-1 and BRCA-2 (Robson, 2000, J. Clin. Oncol. 18:113sup-118sup).

Imaging of breast cancer for diagnosis has been problematic and limited. In addition, dissemination of tumor cells (metastases) to locoregional lymph nodes is an important prognostic factor; five year survival rates drop from 80 percent in patients with no lymph node metastases to 45 to 50 percent in those patients who do have lymph node metastases. A recent report showed that micrometastases can be detected from lymph nodes using reverse transcriptase-PCR methods based on the presence of mRNA for carcinoembryonic antigen, which has previously been shown to be present in the vast majority of breast cancers but not in normal tissues. Liefers et al., New England J. of Med. 339(4):223 (1998).

The identification of novel therapeutic targets and diagnostic markers is essential for improving the current treatment of breast cancer patients. Recent advances in molecular medicine have increased the interest in tumor-specific cell surface antigens that could serve as targets for various immunotherapeutic or small molecule strategies. Antigens suitable for immunotherapeutic strategies should be highly expressed in cancer tissues and ideally not expressed in normal adult tissues. Expression in tissues that are dispensable for life, however, may be tolerated. Examples of such antigens include Her2/neu and the B-cell antigen CD20. Humanized monclonal antibodies directed to Her2/neu (Herceptin®/trastuzumab) are currently in use for the treatment of metastatic breast cancer (Ross and Fletcher, 1998, Stem Cells 16:413-428). Similarly, anti-CD20 monoclonal antibodies (Rituxin®/rituximab) are used to effectively treat non-Hodgekin's lymphoma (Maloney et al., 1997, Blood 90:2188-2195; Leget and Czuczman, 1998, Curr. Opin. Oncol. 10:548-551).

Other potential immunotherapeutic targets have been identified for breast cancer. One such target is polymorphic epithelial mucin (MUC1). MUC1 is a transmembrane protein, present at the apical surface of glandular epithelial cells. It is often overexpressed in breast cancer, and typically exhibits an altered glycosylation pattern, resulting in an antigenically distinct molecule, and is in early clinical trials as a vaccine target (Gilewski et al., 2000, Clin. Cancer Res. 6:1693-1701; Scholl et al., 2000, J. Immunother. 23:570-580). The tumor-expressed protein is often cleaved into the circulation, where it is detectable as the

tumor marker, CA 15-3 (Bon et al., 1997, Clin. Chem. 43:585-593). However, many patients have tumors that express neither HER2 nor MUC-1; therefore, it is clear that other targets need to be identified to manage localized and metastatic disease. Many other genes have been reported to be overexpressed in breast cancer, such as EGFR (Sainsbury et al., 1987, Lancet 1(8547):1398-1402), c-erbB3 (Naidu et al., 1988, Br. J. Cancer 78:1385-1390), FGFR2 (Penault-Llorca et al., 1991, Int. J. Cancer 61:170-176), PKW (Preiherr et al., 2000, Anticancer Res. 20:2255-2264), MTA1 (Nawa et al., 2000, J. Cell Biochem. 79:202-212), breast cancer associated gene 1 (Kurt et al., 2000, Breast Cancer Res. Treat. 59:41-48). Although monoclonal antibodies to the protein products of some of these overexpressed genes have been reported (for review, see Green et al., 2000, Cancer Treat. Rev. 26:269-286), none are currently approved for breast cancer therapy in the US.

Disclosures of certain genes and ESTs described as being expressed in breast cancer are found in international patent applications WO-99/33869, WO-97/25426, WO-97/02280 and WO-00/55173, WO-98/45328 and WO-00/22130. Similarly, genes and ESTs described as being expressed in breast cancer are disclosed in US Patent Nos. 5,759,776 and 5,693,522. The utility of such genes is described in each of these publications, and their disclosures are incorporated herein in their entirety.

While industry and academia have identified novel sequences, there has not been an equal effort exerted to identify the function of these novel sequences. The elucidation of a role for novel proteins and compounds in disease states for identification of therapeutic targets and diagnostic markers is essential for improving the current treatment of breast cancer patients. Accordingly, provided herein are molecular targets for therapeutic intervention in breast and other cancers. Additionally, provided herein are methods that can be used in diagnosis and prognosis of breast cancer. Further provided are methods that can be used to screen candidate bioactive agents for the ability to modulate breast cancer.

SUMMARY OF THE INVENTION

The present invention therefore provides nucleotide sequences of genes that are up- and down-regulated in breast cancer cells. Such genes are useful for diagnostic purposes, and also as targets for screening for therapeutic compounds that modulate breast cancer, such as hormones or antibodies. Other aspects of the invention will become apparent to the skilled artisan by the following description of the invention.

In one aspect, the present invention provides a method of detecting a breast cancer-associated transcript in a cell from a patient, the method comprising contacting a biological sample from the patient with a polynucleotide that selectively hybridizes to a sequence at least 80% identical to a sequence as shown in Table 1.

In one embodiment, the present invention provides a method of determining the level of a breast cancer associated transcript in a cell from a patient.

In one embodiment, the present invention provides a method of detecting a breast cancer-associated transcript in a cell from a patient, the method comprising contacting a biological sample from the patient with a polynucleotide that selectively hybridizes to a sequence at least 80% identical to a sequence as shown in Table 1.

In one embodiment, the polynucleotide selectively hybridizes to a sequence at least 95% identical to a sequence as shown in Table 1.

In one embodiment, the biological sample is a tissue sample. In another embodiment, the biological sample comprises isolated nucleic acids, e.g., mRNA.

In one embodiment, the polynucleotide is labeled, e.g., with a fluorescent label.

In one embodiment, the polynucleotide is immobilized on a solid surface.

In one embodiment, the patient is undergoing a therapeutic regimen to treat breast cancer. In another embodiment, the patient is suspected of having metastatic breast cancer.

In one embodiment, the patient is a human.

In one embodiment, the breast cancer associated transcript is mRNA.

In one embodiment, the method further comprises the step of amplifying nucleic acids before the step of contacting the biological sample with the polynucleotide.

In another aspect, the present invention provides a method of monitoring the efficacy of a therapeutic treatment of breast cancer, the method comprising the steps of: (i) providing a biological sample from a patient undergoing the therapeutic treatment; and (ii) determining the level of a breast cancer-associated transcript in the biological sample by contacting the biological sample with a polynucleotide that selectively hybridizes to a sequence at least 80% identical to a sequence as shown in Table 1, thereby monitoring the efficacy of the therapy. In a further embodiment, the patient has metastatic breast cancer. In a further embodiment, the patient has a drug resistant form of breast cancer.

In one embodiment, the method further comprises the step of: (iii) comparing the level of the breast cancer-associated transcript to a level of the breast cancer-associated transcript in a biological sample from the patient prior to, or earlier in, the therapeutic treatment.

Additionally, provided herein is a method of evaluating the effect of a candidate breast cancer drug comprising administering the drug to a patient and removing a cell sample from the patient. The expression profile of the cell is then determined. This method may further comprise comparing the expression profile to an expression profile of a healthy individual. In a preferred embodiment, said expression profile includes a gene of Table 1.

In one aspect, the present invention provides an isolated nucleic acid molecule consisting of a polynucleotide sequence as shown in Table 1.

In one embodiment, an expression vector or cell comprises the isolated nucleic acid.

In one aspect, the present invention provides an isolated polypeptide which is encoded by a nucleic acid molecule having polynucleotide sequence as shown in Table 1.

In another aspect, the present invention provides an antibody that specifically binds to an isolated polypeptide which is encoded by a nucleic acid molecule having polynucleotide sequence as shown in Table 1.

In one embodiment, the antibody is conjugated to an effector component, e.g., a fluorescent label, a radioisotope or a cytotoxic chemical.

In one embodiment, the antibody is an antibody fragment. In another embodiment, the antibody is humanized.

In one aspect, the present invention provides a method of detecting a breast cancer cell in a biological sample from a patient, the method comprising contacting the biological sample with an antibody as described herein.

In another aspect, the present invention provides a method of detecting antibodies specific to breast cancer in a patient, the method comprising contacting a biological sample from the patient with a polypeptide encoded by a nucleic acid comprising a sequence from Table 1.

In another aspect, the present invention provides a method for identifying a compound that modulates a breast cancer-associated polypeptide, the method comprising the

steps of: (i) contacting the compound with a breast cancer-associated polypeptide, the polypeptide encoded by a polynucleotide that selectively hybridizes to a sequence at least 80% identical to a sequence as shown in Table 1; and (ii) determining the functional effect of the compound upon the polypeptide.

In one embodiment, the functional effect is a physical effect, an enzymatic effect, or a chemical effect.

In one embodiment, the polypeptide is expressed in a eukaryotic host cell or cell membrane. In another embodiment, the polypeptide is recombinant.

In one embodiment, the functional effect is determined by measuring ligand binding to the polypeptide.

In another aspect, the present invention provides a method of inhibiting proliferation of a breast cancer-associated cell to treat breast cancer in a patient, the method comprising the step of administering to the subject a therapeutically effective amount of a compound identified as described herein.

In one embodiment, the compound is an antibody.

In another aspect, the present invention provides a drug screening assay comprising the steps of: (i) administering a test compound to a mammal having breast cancer or to a cell sample isolated therefrom; (ii) comparing the level of gene expression of a polynucleotide that selectively hybridizes to a sequence at least 80% identical to a sequence as shown in Table 1 in a treated cell or mammal with the level of gene expression of the polynucleotide in a control cell sample or mammal, wherein a test compound that modulates the level of expression of the polynucleotide is a candidate for the treatment of breast cancer.

In one embodiment, the control is a mammal with breast cancer or a cell sample therefrom that has not been treated with the test compound. In another embodiment, the control is a normal cell or mammal.

In one embodiment, the test compound is administered in varying amounts or concentrations. In another embodiment, the test compound is administered for varying time periods. In another embodiment, the comparison can occur after addition or removal of the drug candidate.

In one embodiment, the levels of a plurality of polynucleotides that selectively hybridize to a sequence at least 80% identical to a sequence as shown in Table 1 are

individually compared to their respective levels in a control cell sample or mammal. In a preferred embodiment the plurality of polynucleotides is from three to ten.

In another aspect, the present invention provides a method for treating a mammal having breast cancer comprising administering a compound identified by the assay described herein.

In another aspect, the present invention provides a pharmaceutical composition for treating a mammal having breast cancer, the composition comprising a compound identified by the assay described herein and a physiologically acceptable excipient.

In one aspect, the present invention provides a method of screening drug candidates by providing a cell expressing a gene that is up- and down-regulated as in a breast cancer. In one embodiment, a gene is selected from Table 1. The method further includes adding a drug candidate to the cell and determining the effect of the drug candidate on the expression of the expression profile gene.

In one embodiment, the method of screening drug candidates includes comparing the level of expression in the absence of the drug candidate to the level of expression in the presence of the drug candidate, wherein the concentration of the drug candidate can vary when present, and wherein the comparison can occur after addition or removal of the drug candidate. In a preferred embodiment, the cell expresses at least two expression profile genes. The profile genes may show an increase or decrease.

Also provided is a method of evaluating the effect of a candidate breast cancer drug comprising administering the drug to a transgenic animal expressing or over-expressing the breast cancer modulatory protein, or an animal lacking the breast cancer modulatory protein, for example as a result of a gene knockout.

Moreover, provided herein is a biochip comprising one or more nucleic acid segments of Table 1, wherein the biochip comprises fewer than 1000 nucleic acid probes. Preferably, at least two nucleic acid segments are included. More preferably, at least three nucleic acid segments are included.

Furthermore, a method of diagnosing a disorder associated with breast cancer is provided. The method comprises determining the expression of a gene of Table 1 or Table 3, preferably a gene of Table 2, in a first tissue type of a first individual, and comparing the distribution to the expression of the gene from a second normal tissue type from the first

individual or a second unaffected individual. A difference in the expression indicates that the first individual has a disorder associated with breast cancer.

In a further embodiment, the biochip also includes a polynucleotide sequence of a gene that is not up- and down-regulated in breast cancer.

In one embodiment a method for screening for a bioactive agent capable of interfering with the binding of a breast cancer modulating protein (breast cancer modulatory protein) or a fragment thereof and an antibody which binds to said breast cancer modulatory protein or fragment thereof. In a preferred embodiment, the method comprises combining a breast cancer modulatory protein or fragment thereof, a candidate bioactive agent and an antibody which binds to said breast cancer modulatory protein or fragment thereof. The method further includes determining the binding of said breast cancer modulatory protein or fragment thereof and said antibody. Wherein there is a change in binding, an agent is identified as an interfering agent. The interfering agent can be an agonist or an antagonist. Preferably, the agent inhibits breast cancer.

Also provided herein are methods of eliciting an immune response in an individual. In one embodiment a method provided herein comprises administering to an individual a composition comprising a breast cancer modulating protein, or a fragment thereof. In another embodiment, the protein is encoded by a nucleic acid selected from those of Table 1.

Further provided herein are compositions capable of eliciting an immune response in an individual. In one embodiment, a composition provided herein comprises a breast cancer modulating protein, preferably encoded by a nucleic acid of Table 1 or Table 3, more preferably of Table 2, or a fragment thereof, and a pharmaceutically acceptable carrier. In another embodiment, said composition comprises a nucleic acid comprising a sequence encoding a breast cancer modulating protein, preferably selected from the nucleic acids of Table 1, and a pharmaceutically acceptable carrier.

Also provided are methods of neutralizing the effect of a breast cancer protein, or a fragment thereof, comprising contacting an agent specific for said protein with said protein in an amount sufficient to effect neutralization. In another embodiment, the protein is encoded by a nucleic acid selected from those of Table 1.

In another aspect of the invention, a method of treating an individual for breast cancer is provided. In one embodiment, the method comprises administering to said

individual an inhibitor of a breast cancer modulating protein. In another embodiment, the method comprises administering to a patient having breast cancer an antibody to a breast cancer modulating protein conjugated to a therapeutic moiety. Such a therapeutic moiety can be a cytotoxic agent or a radioisotope.

DETAILED DESCRIPTION OF THE INVENTION

In accordance with the objects outlined above, the present invention provides novel methods for diagnosis and prognosis evaluation for breast cancer (PC), including metastatic breast cancer, as well as methods for screening for compositions which modulate breast cancer. Also provided are methods for treating breast cancer.

Tables 1-3 provide unigene cluster identification numbers for the nucleotide sequence of genes that exhibit increased or decreased expression in breast cancer samples. Table 1 also provide an exemplar accession number that provides a nucleotide sequence that is part of the unigene cluster.

Definitions

The term "breast cancer protein" or "breast cancer polynucleotide" or "breast cancer-associated transcript" refers to nucleic acid and polypeptide polymorphic variants, alleles, mutants, and interspecies homologues that: (1) have a nucleotide sequence that has greater than about 60% nucleotide sequence identity, 65%, 70%, 75%, 80%, 85%, 90%, preferably 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% or greater nucleotide sequence identity, preferably over a region of over a region of at least about 25, 50, 100, 200, 500, 1000, or more nucleotides, to a nucleotide sequence of or associated with a gene of Table 1; (2) bind to antibodies, e.g., polyclonal antibodies, raised against an immunogen comprising an amino acid sequence encoded by a nucleotide sequence of or associated with a gene of Table 1, and conservatively modified variants thereof; (3) specifically hybridize under stringent hybridization conditions to a nucleic acid sequence, or the complement thereof of Table 1 and conservatively modified variants thereof or (4) have an amino acid sequence that has greater than about 60% amino acid sequence identity, 65%, 70%, 75%, 80%, 85%, 90%, preferably 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% or greater amino sequence identity, preferably over a region of over a region of at least about 25, 50, 100, 200, 500, 1000, or more amino acid, to an amino acid sequence encoded by a nucleotide

sequence of or associated with a gene of Table 1. A polynucleotide or polypeptide sequence is typically from a mammal including, but not limited to, primate, e.g., human; rodent, e.g., rat, mouse, hamster; cow, pig, horse, sheep, or other mammal. A "breast cancer polypeptide" and a "breast cancer polynucleotide," include both naturally occurring or recombinant forms.

A "full length" breast cancer protein or nucleic acid refers to a breast cancer polypeptide or polynucleotide sequence, or a variant thereof, that contains all of the elements normally contained in one or more naturally occurring, wild type breast cancer polynucleotide or polypeptide sequences. The "full length" may be prior to, or after, various stages of post-translation processing or splicing, including alternative splicing.

"Biological sample" as used herein is a sample of biological tissue or fluid that contains nucleic acids or polypeptides, e.g., of a breast cancer protein, polynucleotide or transcript. Such samples include, but are not limited to, tissue isolated from primates, e.g., humans, or rodents, e.g., mice, and rats. Biological samples may also include sections of tissues such as biopsy and autopsy samples, frozen sections taken for histologic purposes, blood, plasma, serum, sputum, stool, tears, mucus, hair, skin, etc. Biological samples also include explants and primary and/or transformed cell cultures derived from patient tissues. A biological sample is typically obtained from a eukaryotic organism, most preferably a mammal such as a primate e.g., chimpanzee or human; cow; dog; cat; a rodent, e.g., guinea pig, rat, mouse; rabbit; or a bird; reptile; or fish.

"Providing a biological sample" means to obtain a biological sample for use in methods described in this invention. Most often, this will be done by removing a sample of cells from an animal, but can also be accomplished by using previously isolated cells (*e.g.*, isolated by another person, at another time, and/or for another purpose), or by performing the methods of the invention *in vivo*. Archival tissues, having treatment or outcome history, will be particularly useful.

The terms "identical" or percent "identity," in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same (i.e., about 60% identity, preferably 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or higher identity over a specified region, when compared and aligned for maximum correspondence over a comparison window or designated region) as measured using a BLAST or BLAST 2.0 sequence comparison algorithms with default

parameters described below, or by manual alignment and visual inspection (*see, e.g.,* NCBI web site http://www.ncbi.nlm.nih.gov/BLAST/ or the like). Such sequences are then said to be "substantially identical." This definition also refers to, or may be applied to, the compliment of a test sequence. The definition also includes sequences that have deletions and/or additions, as well as those that have substitutions, as well as naturally occurring, e.g., polymorphic or allelic variants, and man-made variants. As described below, the preferred algorithms can account for gaps and the like. Preferably, identity exists over a region that is at least about 25 amino acids or nucleotides in length, or more preferably over a region that is 50-100 amino acids or nucleotides in length.

For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are entered into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. Preferably, default program parameters can be used, or alternative parameters can be designated. The sequence comparison algorithm then calculates the percent sequence identities for the test sequences relative to the reference sequence, based on the program parameters.

A "comparison window", as used herein, includes reference to a segment of one of the number of contiguous positions selected from the group consisting typically of from 20 to 600, usually about 50 to about 200, more usually about 100 to about 150 in which a sequence may be compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned. Methods of alignment of sequences for comparison are well-known in the art. Optimal alignment of sequences for comparison can be conducted, e.g., by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by manual alignment and visual inspection (*see, e.g., Current Protocols in Molecular Biology* (Ausubel *et al.*, eds. 1995 supplement)).

Preferred examples of algorithms that are suitable for determining percent sequence identity and sequence similarity include the BLAST and BLAST 2.0 algorithms,

which are described in Altschul *et al., Nuc. Acids Res.* 25:3389-3402 (1977) and Altschul *et al., J. Mol. Biol.* 215:403-410 (1990). BLAST and BLAST 2.0 are used, with the parameters described herein, to determine percent sequence identity for the nucleic acids and proteins of the invention. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, e.g., for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, M=5, N=-4 and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength of 3, and expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff, *Proc. Natl. Acad. Sci. USA* 89:10915 (1989)) alignments (B) of 50, expectation (E) of 10, M=5, N=-4, and a comparison of both strands.

The BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.,* Karlin & Altschul, *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.2, more

preferably less than about 0.01, and most preferably less than about 0.001. Log values may be large negative numbers, e.g., 5, 10, 20, 30, 40, 40, 70, 90, 110, 150, 170, etc.

An indication that two nucleic acid sequences or polypeptides are substantially identical is that the polypeptide encoded by the first nucleic acid is immunologically cross reactive with the antibodies raised against the polypeptide encoded by the second nucleic acid, as described below. Thus, a polypeptide is typically substantially identical to a second polypeptide, e.g., where the two peptides differ only by conservative substitutions. Another indication that two nucleic acid sequences are substantially identical is that the two molecules or their complements hybridize to each other under stringent conditions, as described below. Yet another indication that two nucleic acid sequences are substantially identical is that the same primers can be used to amplify the sequences.

A "host cell" is a naturally occurring cell or a transformed cell that contains an expression vector and supports the replication or expression of the expression vector. Host cells may be cultured cells, explants, cells *in vivo*, and the like. Host cells may be prokaryotic cells such as *E. coli*, or eukaryotic cells such as yeast, insect, amphibian, or mammalian cells such as CHO, HeLa, and the like (*see, e.g.,* the American Type Culture Collection catalog or web site, www.atcc.org).

The terms "isolated," "purified," or "biologically pure" refer to material that is substantially or essentially free from components that normally accompany it as found in its native state. Purity and homogeneity are typically determined using analytical chemistry techniques such as polyacrylamide gel electrophoresis or high performance liquid chromatography. A protein or nucleic acid that is the predominant species present in a preparation is substantially purified. In particular, an isolated nucleic acid is separated from some open reading frames that naturally flank the gene and encode proteins other than protein encoded by the gene. The term "purified" in some embodiments denotes that a nucleic acid or protein gives rise to essentially one band in an electrophoretic gel. Preferably, it means that the nucleic acid or protein is at least 85% pure, more preferably at least 95% pure, and most preferably at least 99% pure. "Purify" or "purification" in other embodiments means removing at least one contaminant from the composition to be purified. In this sense, purification does not require that the purified compound be homogenous, e.g., 100% pure.

The terms "polypeptide," "peptide" and "protein" are used interchangeably herein to refer to a polymer of amino acid residues. The terms apply to amino acid polymers

in which one or more amino acid residue is an artificial chemical mimetic of a corresponding naturally occurring amino acid, as well as to naturally occurring amino acid polymers, those containing modified residues, and non-naturally occurring amino acid polymer.

The term "amino acid" refers to naturally occurring and synthetic amino acids, as well as amino acid analogs and amino acid mimetics that function similarly to the naturally occurring amino acids. Naturally occurring amino acids are those encoded by the genetic code, as well as those amino acids that are later modified, e.g., hydroxyproline, γ-carboxyglutamate, and O-phosphoserine. Amino acid analogs refers to compounds that have the same basic chemical structure as a naturally occurring amino acid, e.g., an α carbon that is bound to a hydrogen, a carboxyl group, an amino group, and an R group, e.g., homoserine, norleucine, methionine sulfoxide, methionine methyl sulfonium. Such analogs may have modified R groups (e.g., norleucine) or modified peptide backbones, but retain the same basic chemical structure as a naturally occurring amino acid. Amino acid mimetics refers to chemical compounds that have a structure that is different from the general chemical structure of an amino acid, but that functions similarly to a naturally occurring amino acid.

Amino acids may be referred to herein by either their commonly known three letter symbols or by the one-letter symbols recommended by the IUPAC-IUB Biochemical Nomenclature Commission. Nucleotides, likewise, may be referred to by their commonly accepted single-letter codes.

"Conservatively modified variants" applies to both amino acid and nucleic acid sequences. With respect to particular nucleic acid sequences, conservatively modified variants refers to those nucleic acids which encode identical or essentially identical amino acid sequences, or where the nucleic acid does not encode an amino acid sequence, to essentially identical or associated, e.g., naturally contiguous, sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode most proteins. For instance, the codons GCA, GCC, GCG and GCU all encode the amino acid alanine. Thus, at every position where an alanine is specified by a codon, the codon can be altered to another of the corresponding codons described without altering the encoded polypeptide. Such nucleic acid variations are "silent variations," which are one species of conservatively modified variations. Every nucleic acid sequence herein which encodes a polypeptide also describes silent variations of the nucleic acid. One of skill will recognize that in certain contexts each codon in a nucleic acid (except AUG, which is ordinarily the

only codon for methionine, and TGG, which is ordinarily the only codon for tryptophan) can be modified to yield a functionally identical molecule. Accordingly, often silent variations of a nucleic acid which encodes a polypeptide is implicit in a described sequence with respect to the expression product, but not with respect to actual probe sequences.

As to amino acid sequences, one of skill will recognize that individual substitutions, deletions or additions to a nucleic acid, peptide, polypeptide, or protein sequence which alters, adds or deletes a single amino acid or a small percentage of amino acids in the encoded sequence is a "conservatively modified variant" where the alteration results in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. Such conservatively modified variants are in addition to and do not exclude polymorphic variants, interspecies homologs, and alleles of the invention.typically conservative substitutions for one another: 1) Alanine (A), Glycine (G); 2) Aspartic acid (D), Glutamic acid (E); 3) Asparagine (N), Glutamine (Q); 4) Arginine (R), Lysine (K); 5) Isoleucine (I), Leucine (L), Methionine (M), Valine (V); 6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W); 7) Serine (S), Threonine (T); and 8) Cysteine (C), Methionine (M) (*see, e.g.,* Creighton, *Proteins* (1984)).

Macromolecular structures such as polypeptide structures can be described in terms of various levels of organization. For a general discussion of this organization, *see, e.g.,* Alberts *et al.*, *Molecular Biology of the Cell* (3rd ed., 1994) and Cantor & Schimmel, *Biophysical Chemistry Part I: The Conformation of Biological Macromolecules* (1980). "Primary structure" refers to the amino acid sequence of a particular peptide. "Secondary structure" refers to locally ordered, three dimensional structures within a polypeptide. These structures are commonly known as domains. Domains are portions of a polypeptide that often form a compact unit of the polypeptide and are typically 25 to approximately 500 amino acids long. Typical domains are made up of sections of lesser organization such as stretches of β-sheet and α-helices. "Tertiary structure" refers to the complete three dimensional structure of a polypeptide monomer. "Quaternary structure" refers to the three dimensional structure formed, usually by the noncovalent association of independent tertiary units. Anisotropic terms are also known as energy terms.

"Nucleic acid" or "oligonucleotide" or "polynucleotide" or grammatical equivalents used herein means at least two nucleotides covalently linked together.

Oligonucleotides are typically from about 5, 6, 7, 8, 9, 10, 12, 15, 25, 30, 40, 50 or more nucleotides in length, up to about 100 nucleotides in length. Nucleic acids and polynucleotides are a polymers of any length, including longer lengths, e.g., 200, 300, 500, 1000, 2000, 3000, 5000, 7000, 10,000, etc. A nucleic acid of the present invention will generally contain phosphodiester bonds, although in some cases, nucleic acid analogs are included that may have alternate backbones, comprising, e.g., phosphoramidate, phosphorothioate, phosphorodithioate, or O-methylphophoroamidite linkages (see Eckstein, Oligonucleotides and Analogues: A Practical Approach, Oxford University Press); and peptide nucleic acid backbones and linkages. Other analog nucleic acids include those with positive backbones; non-ionic backbones, and non-ribose backbones, including those described in U.S. Patent Nos. 5,235,033 and 5,034,506, and Chapters 6 and 7, ASC Symposium Series 580, *Carbohydrate Modifications in Antisense Research*, Sanghui & Cook, eds.. Nucleic acids containing one or more carbocyclic sugars are also included within one definition of nucleic acids. Modifications of the ribose-phosphate backbone may be done for a variety of reasons, e.g. to increase the stability and half-life of such molecules in physiological environments or as probes on a biochip. Mixtures of naturally occurring nucleic acids and analogs can be made; alternatively, mixtures of different nucleic acid analogs, and mixtures of naturally occurring nucleic acids and analogs may be made.

A variety of references disclose such nucleic acid analogs, including, for example, phosphoramidate (Beaucage et al., Tetrahedron 49(10):1925 (1993) and references therein; Letsinger, J. Org. Chem. 35:3800 (1970); Sprinzl et al., Eur. J. Biochem. 81:579 (1977); Letsinger et al., Nucl. Acids Res. 14:3487 (1986); Sawai et al, Chem. Lett. 805 (1984), Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); and Pauwels et al., Chemica Scripta 26:141 91986)), phosphorothioate (Mag et al., Nucleic Acids Res. 19:1437 (1991); and U.S. Patent No. 5,644,048), phosphorodithioate (Briu et al., J. Am. Chem. Soc. 111:2321 (1989), O-methylphophoroamidite linkages (see Eckstein, Oligonucleotides and Analogues: A Practical Approach, Oxford University Press), and peptide nucleic acid backbones and linkages (see Egholm, J. Am. Chem. Soc. 114:1895 (1992); Meier et al., Chem. Int. Ed. Engl. 31:1008 (1992); Nielsen, Nature, 365:566 (1993); Carlsson et al., Nature 380:207 (1996), all of which are incorporated by reference). Other analog nucleic acids include those with positive backbones (Denpcy et al., Proc. Natl. Acad. Sci. USA 92:6097 (1995); non-ionic backbones (U.S. Patent Nos. 5,386,023, 5,637,684, 5,602,240, 5,216,141 and 4,469,863;

Kiedrowshi et al., Angew. Chem. Intl. Ed. English 30:423 (1991); Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); Letsinger et al., Nucleoside & Nucleotide 13:1597 (1994); Chapters 2 and 3, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y.S. Sanghui and P. Dan Cook; Mesmaeker et al., Bioorganic & Medicinal Chem. Lett. 4:395 (1994); Jeffs et al., J. Biomolecular NMR 34:17 (1994); Tetrahedron Lett. 37:743 (1996)) and non-ribose backbones, including those described in U.S. Patent Nos. 5,235,033 and 5,034,506, and Chapters 6 and 7, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y.S. Sanghui and P. Dan Cook. Nucleic acids containing one or more carbocyclic sugars are also included within one definition of nucleic acids (see Jenkins et al., Chem. Soc. Rev. (1995) pp 169-176). Several nucleic acid analogs are described in Rawls, C & E News June 2, 1997 page 35. All of these references are hereby expressly incorporated by reference.

Particularly preferred are peptide nucleic acids (PNA) which includes peptide nucleic acid analogs. These backbones are substantially non-ionic under neutral conditions, in contrast to the highly charged phosphodiester backbone of naturally occurring nucleic acids. This results in two advantages. First, the PNA backbone exhibits improved hybridization kinetics. PNAs have larger changes in the melting temperature (T_m) for mismatched versus perfectly matched basepairs. DNA and RNA typically exhibit a 2-4°C drop in T_m for an internal mismatch. With the non-ionic PNA backbone, the drop is closer to 7-9°C. Similarly, due to their non-ionic nature, hybridization of the bases attached to these backbones is relatively insensitive to salt concentration. In addition, PNAs are not degraded by cellular enzymes, and thus can be more stable.

The nucleic acids may be single stranded or double stranded, as specified, or contain portions of both double stranded or single stranded sequence. As will be appreciated by those in the art, the depiction of a single strand also defines the sequence of the complementary strand; thus the sequences described herein also provide the complement of the sequence. The nucleic acid may be DNA, both genomic and cDNA, RNA or a hybrid, where the nucleic acid may contain combinations of deoxyribo- and ribo-nucleotides, and combinations of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xanthine hypoxanthine, isocytosine, isoguanine, etc. "Transcript" typically refers to a naturally occurring RNA, e.g., a pre-mRNA, hnRNA, or mRNA. As used herein, the term "nucleoside" includes nucleotides and nucleoside and nucleotide analogs, and modified

nucleosides such as amino modified nucleosides. In addition, "nucleoside" includes non-naturally occurring analog structures. Thus, e.g. the individual units of a peptide nucleic acid, each containing a base, are referred to herein as a nucleoside.

A "label" or a "detectable moiety" is a composition detectable by spectroscopic, photochemical, biochemical, immunochemical, chemical, or other physical means. For example, useful labels include ^{32}P, fluorescent dyes, electron-dense reagents, enzymes (*e.g.*, as commonly used in an ELISA), biotin, digoxigenin, or haptens and proteins or other entities which can be made detectable, *e.g.*, by incorporating a radiolabel into the peptide or used to detect antibodies specifically reactive with the peptide. The labels may be incorporated into the breast cancer nucleic acids, proteins and antibodies at any position. Any method known in the art for conjugating the antibody to the label may be employed, including those methods described by Hunter et al., Nature, 144:945 (1962); David et al., Biochemistry, 13:1014 (1974); Pain et al., J. Immunol. Meth., 40:219 (1981); and Nygren, J. Histochem. and Cytochem., 30:407 (1982).

An "effector" or "effector moiety" or "effector component" is a molecule that is bound (or linked, or conjugated), either covalently, through a linker or a chemical bond, or noncovalently, through ionic, van der Waals, electrostatic, or hydrogen bonds, to an antibody. The "effector" can be a variety of molecules including, e.g., detection moieties including radioactive compounds, fluorescent compounds, an enzyme or substrate, tags such as epitope tags, a toxin; activatable moieties, a chemotherapeutic agent; a lipase; an antibiotic; or a radioisotope emitting "hard" e.g., beta radiation.

A "labeled nucleic acid probe or oligonucleotide" is one that is bound, either covalently, through a linker or a chemical bond, or noncovalently, through ionic, van der Waals, electrostatic, or hydrogen bonds to a label such that the presence of the probe may be detected by detecting the presence of the label bound to the probe. Alternatively, method using high affinity interactions may achieve the same results where one of a pair of binding partners binds to the other, e.g., biotin, streptavidin.

As used herein a "nucleic acid probe or oligonucleotide" is defined as a nucleic acid capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, usually through hydrogen bond formation. As used herein, a probe may include natural (i.e., A, G, C, or T) or modified bases (7-deazaguanosine, inosine, etc.). In addition, the bases in a probe

may be joined by a linkage other than a phosphodiester bond, so long as it does not functionally interfere with hybridization. Thus, e.g., probes may be peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than phosphodiester linkages. It will be understood by one of skill in the art that probes may bind target sequences lacking complete complementarity with the probe sequence depending upon the stringency of the hybridization conditions. The probes are preferably directly labeled as with isotopes, chromophores, lumiphores, chromogens, or indirectly labeled such as with biotin to which a streptavidin complex may later bind. By assaying for the presence or absence of the probe, one can detect the presence or absence of the select sequence or subsequence. Diagnosis or prognosis may be based at the genomic level, or at the level of RNA or protein expression.

The term "recombinant" when used with reference, e.g., to a cell, or nucleic acid, protein, or vector, indicates that the cell, nucleic acid, protein or vector, has been modified by the introduction of a heterologous nucleic acid or protein or the alteration of a native nucleic acid or protein, or that the cell is derived from a cell so modified. Thus, e.g., recombinant cells express genes that are not found within the native (non-recombinant) form of the cell or express native genes that are otherwise abnormally expressed, under expressed or not expressed at all. By the term "recombinant nucleic acid" herein is meant nucleic acid, originally formed *in vitro*, in general, by the manipulation of nucleic acid, e.g., using polymerases and endonucleases, in a form not normally found in nature. In this manner, operably linkage of different sequences is achieved. Thus an isolated nucleic acid, in a linear form, or an expression vector formed *in vitro* by ligating DNA molecules that are not normally joined, are both considered recombinant for the purposes of this invention. It is understood that once a recombinant nucleic acid is made and reintroduced into a host cell or organism, it will replicate non-recombinantly, i.e., using the *in vivo* cellular machinery of the host cell rather than *in vitro* manipulations; however, such nucleic acids, once produced recombinantly, although subsequently replicated non-recombinantly, are still considered recombinant for the purposes of the invention. Similarly, a "recombinant protein" is a protein made using recombinant techniques, i.e., through the expression of a recombinant nucleic acid as depicted above.

The term "heterologous" when used with reference to portions of a nucleic acid indicates that the nucleic acid comprises two or more subsequences that are not normally found in the same relationship to each other in nature. For instance, the nucleic acid is

typically recombinantly produced, having two or more sequences, e.g., from unrelated genes arranged to make a new functional nucleic acid, e.g., a promoter from one source and a coding region from another source. Similarly, a heterologous protein will often refer to two or more subsequences that are not found in the same relationship to each other in nature (e.g., a fusion protein).

A "promoter" is defined as an array of nucleic acid control sequences that direct transcription of a nucleic acid. As used herein, a promoter includes necessary nucleic acid sequences near the start site of transcription, such as, in the case of a polymerase II type promoter, a TATA element. A promoter also optionally includes distal enhancer or repressor elements, which can be located as much as several thousand base pairs from the start site of transcription. A "constitutive" promoter is a promoter that is active under most environmental and developmental conditions. An "inducible" promoter is a promoter that is active under environmental or developmental regulation. The term "operably linked" refers to a functional linkage between a nucleic acid expression control sequence (such as a promoter, or array of transcription factor binding sites) and a second nucleic acid sequence, wherein the expression control sequence directs transcription of the nucleic acid corresponding to the second sequence.

An "expression vector" is a nucleic acid construct, generated recombinantly or synthetically, with a series of specified nucleic acid elements that permit transcription of a particular nucleic acid in a host cell. The expression vector can be part of a plasmid, virus, or nucleic acid fragment. Typically, the expression vector includes a nucleic acid to be transcribed operably linked to a promoter.

The phrase "selectively (or specifically) hybridizes to" refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent hybridization conditions when that sequence is present in a complex mixture (e.g., total cellular or library DNA or RNA).

The phrase "stringent hybridization conditions" refers to conditions under which a probe will hybridize to its target subsequence, typically in a complex mixture of nucleic acids, but to no other sequences. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen, *Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic*

Probes, "Overview of principles of hybridization and the strategy of nucleic acid assays" (1993). Generally, stringent conditions are selected to be about 5-10°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength pH. The T_m is the temperature (under defined ionic strength, pH, and nucleic concentration) at which 50% of the probes complementary to the target hybridize to the target sequence at equilibrium (as the target sequences are present in excess, at T_m, 50% of the probes are occupied at equilibrium). Stringent conditions will be those in which the salt concentration is less than about 1.0 M sodium ion, typically about 0.01 to 1.0 M sodium ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (e.g., 10 to 50 nucleotides) and at least about 60°C for long probes (e.g., greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide. For selective or specific hybridization, a positive signal is at least two times background, preferably 10 times background hybridization. Exemplary stringent hybridization conditions can be as following: 50% formamide, 5x SSC, and 1% SDS, incubating at 42°C, or, 5x SSC, 1% SDS, incubating at 65°C, with wash in 0.2x SSC, and 0.1% SDS at 65°C. For PCR, a temperature of about 36°C is typical for low stringency amplification, although annealing temperatures may vary between about 32°C and 48°C depending on primer length. For high stringency PCR amplification, a temperature of about 62°C is typical, although high stringency annealing temperatures can range from about 50°C to about 65°C, depending on the primer length and specificity. Typical cycle conditions for both high and low stringency amplifications include a denaturation phase of 90°C - 95°C for 30 sec - 2 min., an annealing phase lasting 30 sec. - 2 min., and an extension phase of about 72°C for 1 - 2 min. Protocols and guidelines for low and high stringency amplification reactions are provided, e.g., in Innis *et al.* (1990) *PCR Protocols, A Guide to Methods and Applications*, Academic Press, Inc. N.Y.).

Nucleic acids that do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides which they encode are substantially identical. This occurs, e.g., when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code. In such cases, the nucleic acids typically hybridize under moderately stringent hybridization conditions. Exemplary "moderately stringent hybridization conditions" include a hybridization in a buffer of 40% formamide, 1 M NaCl,

1% SDS at 37°C, and a wash in 1X SSC at 45°C. A positive hybridization is at least twice background. Those of ordinary skill will readily recognize that alternative hybridization and wash conditions can be utilized to provide conditions of similar stringency. Additional guidelines for determining hybridization parameters are provided in numerous reference, e.g., and Current Protocols in Molecular Biology, ed. Ausubel, *et al.*

The phrase "functional effects" in the context of assays for testing compounds that modulate activity of a breast cancer protein includes the determination of a parameter that is indirectly or directly under the influence of the breast cancer protein or nucleic acid, e.g., a functional, physical, or chemical effect, such as the ability to decrease breast cancer. It includes ligand binding activity; cell growth on soft agar; anchorage dependence; contact inhibition and density limitation of growth; cellular proliferation; cellular transformation; growth factor or serum dependence; tumor specific marker levels; invasiveness into Matrigel; tumor growth and metastasis *in vivo*; mRNA and protein expression in cells undergoing metastasis, and other characteristics of breast cancer cells. "Functional effects" include *in vitro*, *in vivo*, and *ex vivo* activities.

By "determining the functional effect" is meant assaying for a compound that increases or decreases a parameter that is indirectly or directly under the influence of a breast cancer protein sequence, e.g., functional, enzymatic, physical and chemical effects. Such functional effects can be measured by any means known to those skilled in the art, e.g., changes in spectroscopic characteristics (e.g., fluorescence, absorbance, refractive index), hydrodynamic (e.g., shape), chromatographic, or solubility properties for the protein, measuring inducible markers or transcriptional activation of the breast cancer protein; measuring binding activity or binding assays, e.g. binding to antibodies or other ligands, and measuring cellular proliferation. Determination of the functional effect of a compound on breast cancer can also be performed using breast cancer assays known to those of skill in the art such as an *in vitro* assays, e.g., cell growth on soft agar; anchorage dependence; contact inhibition and density limitation of growth; cellular proliferation; cellular transformation; growth factor or serum dependence; tumor specific marker levels; invasiveness into Matrigel; tumor growth and metastasis *in vivo*; mRNA and protein expression in cells undergoing metastasis, and other characteristics of breast cancer cells. The functional effects can be evaluated by many means known to those skilled in the art, e.g., microscopy for quantitative or qualitative measures of alterations in morphological features, measurement of changes in

RNA or protein levels for breast cancer-associated sequences, measurement of RNA stability, identification of downstream or reporter gene expression (CAT, luciferase, β-gal, GFP and the like), e.g., via chemiluminescence, fluorescence, colorimetric reactions, antibody binding, inducible markers, and ligand binding assays.

"Inhibitors", "activators", and "modulators" of breast cancer polynucleotide and polypeptide sequences are used to refer to activating, inhibitory, or modulating molecules or compounds identified using *in vitro* and *in vivo* assays of breast cancer polynucleotide and polypeptide sequences. Inhibitors are compounds that, e.g., bind to, partially or totally block activity, decrease, prevent, delay activation, inactivate, desensitize, or down regulate the activity or expression of breast cancer proteins, e.g., antagonists. Antisense nucleic acids may seem to inhibit expression and subsequent function of the protein. "Activators" are compounds that increase, open, activate, facilitate, enhance activation, sensitize, agonize, or up regulate breast cancer protein activity. Inhibitors, activators, or modulators also include genetically modified versions of breast cancer proteins, e.g., versions with altered activity, as well as naturally occurring and synthetic ligands, antagonists, agonists, antibodies, small chemical molecules and the like. Such assays for inhibitors and activators include, e.g., expressing the breast cancer protein *in vitro*, in cells, or cell membranes, applying putative modulator compounds, and then determining the functional effects on activity, as described above. Activators and inhibitors of breast cancer can also be identified by incubating breast cancer cells with the test compound and determining increases or decreases in the expression of 1 or more breast cancer proteins, e.g., 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 40, 50 or more breast cancer proteins, such as breast cancer proteins encoded by the sequences set out in Table 1.

Samples or assays comprising breast cancer proteins that are treated with a potential activator, inhibitor, or modulator are compared to control samples without the inhibitor, activator, or modulator to examine the extent of inhibition. Control samples (untreated with inhibitors) are assigned a relative protein activity value of 100%. Inhibition of a polypeptide is achieved when the activity value relative to the control is about 80%, preferably 50%, more preferably 25-0%. Activation of a breast cancer polypeptide is achieved when the activity value relative to the control (untreated with activators) is 110%, more preferably 150%, more preferably 200-500% (i.e., two to five fold higher relative to the control), more preferably 1000-3000% higher.

The phrase "changes in cell growth" refers to any change in cell growth and proliferation characteristics *in vitro* or *in vivo*, such as formation of foci, anchorage independence, semi-solid or soft agar growth, changes in contact inhibition and density limitation of growth, loss of growth factor or serum requirements, changes in cell morphology, gaining or losing immortalization, gaining or losing tumor specific markers, ability to form or suppress tumors when injected into suitable animal hosts, and/or immortalization of the cell. *See, e.g.,* Freshney, *Culture of Animal Cells a Manual of Basic Technique* pp. 231-241 (3rd ed. 1994).

"Tumor cell" refers to precancerous, cancerous, and normal cells in a tumor.

"Cancer cells," "transformed" cells or "transformation" in tissue culture, refers to spontaneous or induced phenotypic changes that do not necessarily involve the uptake of new genetic material. Although transformation can arise from infection with a transforming virus and incorporation of new genomic DNA, or uptake of exogenous DNA, it can also arise spontaneously or following exposure to a carcinogen, thereby mutating an endogenous gene. Transformation is associated with phenotypic changes, such as immortalization of cells, aberrant growth control, nonmorphological changes, and/or malignancy (*see*, Freshney, *Culture of Animal Cells a Manual of Basic Technique* (3rd ed. 1994)).

"Antibody" refers to a polypeptide comprising a framework region from an immunoglobulin gene or fragments thereof that specifically binds and recognizes an antigen. The recognized immunoglobulin genes include the kappa, lambda, alpha, gamma, delta, epsilon, and mu constant region genes, as well as the myriad immunoglobulin variable region genes. Light chains are classified as either kappa or lambda. Heavy chains are classified as gamma, mu, alpha, delta, or epsilon, which in turn define the immunoglobulin classes, IgG, IgM, IgA, IgD and IgE, respectively. Typically, the antigen-binding region of an antibody or its functional equivalent will be most critical in specificity and affinity of binding. See Paul, *Fundamental Immunology*.

An exemplary immunoglobulin (antibody) structural unit comprises a tetramer. Each tetramer is composed of two identical pairs of polypeptide chains, each pair having one "light" (about 25 kD) and one "heavy" chain (about 50-70 kD). The N-terminus of each chain defines a variable region of about 100 to 110 or more amino acids primarily responsible for antigen recognition. The terms variable light chain (V_L) and variable heavy chain (V_H) refer to these light and heavy chains respectively.

Antibodies exist, e.g., as intact immunoglobulins or as a number of well-characterized fragments produced by digestion with various peptidases. Thus, e.g., pepsin digests an antibody below the disulfide linkages in the hinge region to produce F(ab)'$_2$, a dimer of Fab which itself is a light chain joined to V_H-C_H1 by a disulfide bond. The F(ab)'$_2$ may be reduced under mild conditions to break the disulfide linkage in the hinge region, thereby converting the F(ab)'$_2$ dimer into an Fab' monomer. The Fab' monomer is essentially Fab with part of the hinge region (*see Fundamental Immunology* (Paul ed., 3d ed. 1993). While various antibody fragments are defined in terms of the digestion of an intact antibody, one of skill will appreciate that such fragments may be synthesized *de novo* either chemically or by using recombinant DNA methodology. Thus, the term antibody, as used herein, also includes antibody fragments either produced by the modification of whole antibodies, or those synthesized *de novo* using recombinant DNA methodologies (e.g., single chain Fv) or those identified using phage display libraries (*see, e.g.,* McCafferty *et al.*, *Nature* 348:552-554 (1990))

For preparation of antibodies, e.g., recombinant, monoclonal, or polyclonal antibodies, many technique known in the art can be used (*see, e.g.,* Kohler & Milstein, *Nature* 256:495-497 (1975); Kozbor *et al.*, *Immunology Today* 4:72 (1983); Cole *et al.*, pp. 77-96 in *Monoclonal Antibodies and Cancer Therapy* (1985); Coligan, *Current Protocols in Immunology* (1991); Harlow & Lane, *Antibodies, A Laboratory Manual* (1988); and Goding, *Monoclonal Antibodies: Principles and Practice* (2d ed. 1986)). Techniques for the production of single chain antibodies (U.S. Patent 4,946,778) can be adapted to produce antibodies to polypeptides of this invention. Also, transgenic mice, or other organisms such as other mammals, may be used to express humanized antibodies. Alternatively, phage display technology can be used to identify antibodies and heteromeric Fab fragments that specifically bind to selected antigens (*see, e.g.,* McCafferty *et al.*, *Nature* 348:552-554 (1990); Marks *et al.*, *Biotechnology* 10:779-783 (1992)).

A "chimeric antibody" is an antibody molecule in which (a) the constant region, or a portion thereof, is altered, replaced or exchanged so that the antigen binding site (variable region) is linked to a constant region of a different or altered class, effector function and/or species, or an entirely different molecule which confers new properties to the chimeric antibody, *e.g.*, an enzyme, toxin, hormone, growth factor, drug, etc.; or (b) the variable

region, or a portion thereof, is altered, replaced or exchanged with a variable region having a different or altered antigen specificity.

Identification of breast cancer-associated sequences

In one aspect, the expression levels of genes are determined in different patient samples for which diagnosis information is desired, to provide expression profiles. An expression profile of a particular sample is essentially a "fingerprint" of the state of the sample; while two states may have any particular gene similarly expressed, the evaluation of a number of genes simultaneously allows the generation of a gene expression profile that is characteristic of the state of the cell. That is, normal tissue (e.g., normal breast or other tissue) may be distinguished from cancerous or metastatic cancerous tissue of the breast, or breast cancer tissue or metastatic breast cancerous tissue can be compared with tissue samples of breast and other tissues from surviving cancer patients. By comparing expression profiles of tissue in known different breast cancer states, information regarding which genes are important (including both up- and down-regulation of genes) in each of these states is obtained.

The identification of sequences that are differentially expressed in breast cancer versus non-breast cancer tissue allows the use of this information in a number of ways. For example, a particular treatment regime may be evaluated: does a chemotherapeutic drug act to down-regulate breast cancer, and thus tumor growth or recurrence, in a particular patient. Similarly, diagnosis and treatment outcomes may be done or confirmed by comparing patient samples with the known expression profiles. Metastatic tissue can also be analyzed to determine the stage of breast cancer in the tissue. Furthermore, these gene expression profiles (or individual genes) allow screening of drug candidates with an eye to mimicking or altering a particular expression profile; e.g., screening can be done for drugs that suppress the breast cancer expression profile. This may be done by making biochips comprising sets of the important breast cancer genes, which can then be used in these screens. These methods can also be done on the protein basis; that is, protein expression levels of the breast cancer proteins can be evaluated for diagnostic purposes or to screen candidate agents. In addition, the breast cancer nucleic acid sequences can be administered for gene therapy purposes, including the administration of antisense nucleic acids, or the breast cancer proteins (including antibodies and other modulators thereof) administered as therapeutic drugs.

Thus the present invention provides nucleic acid and protein sequences that are differentially expressed in breast cancer, herein termed "breast cancer sequences." As outlined below, breast cancer sequences include those that are up-regulated (i.e., expressed at a higher level) in breast cancer, as well as those that are down-regulated (i.e., expressed at a lower level). In a preferred embodiment, the breast cancer sequences are from humans; however, as will be appreciated by those in the art, breast cancer sequences from other organisms may be useful in animal models of disease and drug evaluation; thus, other breast cancer sequences are provided, from vertebrates, including mammals, including rodents (rats, mice, hamsters, guinea pigs, etc.), primates, farm animals (including sheep, goats, pigs, cows, horses, etc.) and pets, e.g., (dogs, cats, etc.). Breast cancer sequences from other organisms may be obtained using the techniques outlined below.

Breast cancer sequences can include both nucleic acid and amino acid sequences. As will be appreciated by those in the art and is more fully outlined below, breast cancer nucleic acid sequences are useful in a variety of applications, including diagnostic applications, which will detect naturally occurring nucleic acids, as well as screening applications; e.g., biochips comprising nucleic acid probes or PCR microtiter plates with selected probes to the breast cancer sequences can be generated.

A breast cancer sequence can be initially identified by substantial nucleic acid and/or amino acid sequence homology to the breast cancer sequences outlined herein. Such homology can be based upon the overall nucleic acid or amino acid sequence, and is generally determined as outlined below, using either homology programs or hybridization conditions.

For identifying breast cancer-associated sequences, the breast cancer screen typically includes comparing genes identified in different tissues, e.g., normal and cancerous tissues, or tumor tissue samples from patients who have metastatic disease vs. non metastatic tissue. Other suitable tissue comparisons include comparing breast cancer samples with metastatic cancer samples from other cancers, such as lung, breast, gastrointestinal cancers, ovarian, etc. Samples of different stages of breast cancer, e.g., survivor tissue, drug resistant states, and tissue undergoing metastasis, are applied to biochips comprising nucleic acid probes. The samples are first microdissected, if applicable, and treated as is known in the art for the preparation of mRNA. Suitable biochips are commercially available, e.g. from

Affymetrix. Gene expression profiles as described herein are generated and the data analyzed.

In one embodiment, the genes showing changes in expression as between normal and disease states are compared to genes expressed in other normal tissues, preferably normal breast, but also including, and not limited to lung, heart, brain, liver, breast, kidney, muscle, colon, small intestine, large intestine, spleen, bone and placenta. In a preferred embodiment, those genes identified during the breast cancer screen that are expressed in any significant amount in other tissues are removed from the profile, although in some embodiments, this is not necessary. That is, when screening for drugs, it is usually preferable that the target be disease specific, to minimize possible side effects.

In a preferred embodiment, breast cancer sequences are those that are up-regulated in breast cancer; that is, the expression of these genes is higher in the breast cancer tissue as compared to non-cancerous tissue. "Up-regulation" as used herein often means at least about a two-fold change, preferably at least about a three fold change, with at least about five-fold or higher being preferred. All unigene cluster identification numbers and accession numbers herein are for the GenBank sequence database and the sequences of the accession numbers are hereby expressly incorporated by reference. GenBank is known in the art, *see, e.g.,* Benson, DA, *et al.*, Nucleic Acids Research 26:1-7 (1998) and http://www.ncbi.nlm.nih.gov/. Sequences are also available in other databases, e.g., European Molecular Biology Laboratory (EMBL) and DNA Database of Japan (DDBJ). U.S. Patent Application N. 09/687,576, with the same assignee as the present application, further discloses related sequences, compositions, and methods of diagnosis and treatment of breast cancer is hereby expressly incorporated by reference.

In another preferred embodiment, breast cancer sequences are those that are down-regulated in the breast cancer; that is, the expression of these genes is lower in breast cancer tissue as compared to non-cancerous tissue (*see, e.g.,* Tables 1 and 2). "Down-regulation" as used herein often means at least about a two-fold change, preferably at least about a three fold change, with at least about five-fold or higher being preferred.

Informatics

The ability to identify genes that are over or under expressed in breast cancer can additionally provide high-resolution, high-sensitivity datasets which can be used in the areas of diagnostics, therapeutics, drug development, pharmacogenetics, protein structure, biosensor development, and other related areas. For example, the expression profiles can be used in diagnostic or prognostic evaluation of patients with breast cancer. Or as another example, subcellular toxicological information can be generated to better direct drug structure and activity correlation (*see* Anderson, *Pharmaceutical Proteomics: Targets, Mechanism, and Function*, paper presented at the IBC Proteomics conference, Coronado, CA (June 11-12, 1998)). Subcellular toxicological information can also be utilized in a biological sensor device to predict the likely toxicological effect of chemical exposures and likely tolerable exposure thresholds (*see* U.S. Patent No. 5,811,231). Similar advantages accrue from datasets relevant to other biomolecules and bioactive agents (e.g., nucleic acids, saccharides, lipids, drugs, and the like).

Thus, in another embodiment, the present invention provides a database that includes at least one set of assay data. The data contained in the database is acquired, e.g., using array analysis either singly or in a library format. The database can be in substantially any form in which data can be maintained and transmitted, but is preferably an electronic database. The electronic database of the invention can be maintained on any electronic device allowing for the storage of and access to the database, such as a personal computer, but is preferably distributed on a wide area network, such as the World Wide Web.

The focus of the present section on databases that include peptide sequence data is for clarity of illustration only. It will be apparent to those of skill in the art that similar databases can be assembled for any assay data acquired using an assay of the invention.

The compositions and methods for identifying and/or quantitating the relative and/or absolute abundance of a variety of molecular and macromolecular species from a biological sample undergoing breast cancer, i.e., the identification of breast cancer-associated sequences described herein, provide an abundance of information, which can be correlated with pathological conditions, predisposition to disease, drug testing, therapeutic monitoring, gene-disease causal linkages, identification of correlates of immunity and physiological status, among others. Although the data generated from the assays of the invention is suited

for manual review and analysis, in a preferred embodiment, prior data processing using high-speed computers is utilized.

An array of methods for indexing and retrieving biomolecular information is known in the art. For example, U.S. Patents 6,023,659 and 5,966,712 disclose a relational database system for storing biomolecular sequence information in a manner that allows sequences to be catalogued and searched according to one or more protein function hierarchies. U.S. Patent 5,953,727 discloses a relational database having sequence records containing information in a format that allows a collection of partial-length DNA sequences to be catalogued and searched according to association with one or more sequencing projects for obtaining full-length sequences from the collection of partial length sequences. U.S. Patent 5,706,498 discloses a gene database retrieval system for making a retrieval of a gene sequence similar to a sequence data item in a gene database based on the degree of similarity between a key sequence and a target sequence. U.S. Patent 5,538,897 discloses a method using mass spectroscopy fragmentation patterns of peptides to identify amino acid sequences in computer databases by comparison of predicted mass spectra with experimentally-derived mass spectra using a closeness-of-fit measure. U.S. Patent 5,926,818 discloses a multi-dimensional database comprising a functionality for multi-dimensional data analysis described as on-line analytical processing (OLAP), which entails the consolidation of projected and actual data according to more than one consolidation path or dimension. U.S. Patent 5,295,261 reports a hybrid database structure in which the fields of each database record are divided into two classes, navigational and informational data, with navigational fields stored in a hierarchical topological map which can be viewed as a tree structure or as the merger of two or more such tree structures.

See also Mount *et al., Bioinformatics* (2001); *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Durbin *et al.*, eds., 1999); *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (Baxevanis & Oeullette eds., 1998)); Rashidi & Buehler, *Bioinformatics: Basic Applications in Biological Science and Medicine* (1999); *Introduction to Computational Molecular Biology* (Setubal *et al.*, eds 1997); *Bioinformatics: Methods and Protocols* (Misener & Krawetz, eds, 2000); *Bioinformatics: Sequence, Structure, and Databanks: A Practical Approach* (Higgins & Taylor, eds., 2000); Brown, *Bioinformatics: A Biologist's Guide to Biocomputing and the*

Internet (2001); Han & Kamber, *Data Mining: Concepts and Techniques* (2000); and Waterman, *Introduction to Computational Biology: Maps, Sequences, and Genomes* (1995).

The present invention provides a computer database comprising a computer and software for storing in computer-retrievable form assay data records cross-tabulated, e.g., with data specifying the source of the target-containing sample from which each sequence specificity record was obtained.

In an exemplary embodiment, at least one of the sources of target-containing sample is from a control tissue sample known to be free of pathological disorders. In a variation, at least one of the sources is a known pathological tissue specimen, e.g., a neoplastic lesion or another tissue specimen to be analyzed for breast cancer. In another variation, the assay records cross-tabulate one or more of the following parameters for each target species in a sample: (1) a unique identification code, which can include, e.g., a target molecular structure and/or characteristic separation coordinate (e.g., electrophoretic coordinates); (2) sample source; and (3) absolute and/or relative quantity of the target species present in the sample.

The invention also provides for the storage and retrieval of a collection of target data in a computer data storage apparatus, which can include magnetic disks, optical disks, magneto-optical disks, DRAM, SRAM, SGRAM, SDRAM, RDRAM, DDR RAM, magnetic bubble memory devices, and other data storage devices, including CPU registers and on-CPU data storage arrays. Typically, the target data records are stored as a bit pattern in an array of magnetic domains on a magnetizable medium or as an array of charge states or transistor gate states, such as an array of cells in a DRAM device (e.g., each cell comprised of a transistor and a charge storage area, which may be on the transistor). In one embodiment, the invention provides such storage devices, and computer systems built therewith, comprising a bit pattern encoding a protein expression fingerprint record comprising unique identifiers for at least 10 target data records cross-tabulated with target source.

When the target is a peptide or nucleic acid, the invention preferably provides a method for identifying related peptide or nucleic acid sequences, comprising performing a computerized comparison between a peptide or nucleic acid sequence assay record stored in or retrieved from a computer storage device or database and at least one other sequence. The comparison can include a sequence analysis or comparison algorithm or computer program embodiment thereof (e.g., FASTA, TFASTA, GAP, BESTFIT) and/or the comparison may

be of the relative amount of a peptide or nucleic acid sequence in a pool of sequences determined from a polypeptide or nucleic acid sample of a specimen.

The invention also preferably provides a magnetic disk, such as an IBM-compatible (DOS, Windows, Windows95/98/2000, Windows NT, OS/2) or other format (*e.g.*, Linux, SunOS, Solaris, AIX, SCO Unix, VMS, MV, Macintosh, etc.) floppy diskette or hard (fixed, Winchester) disk drive, comprising a bit pattern encoding data from an assay of the invention in a file format suitable for retrieval and processing in a computerized sequence analysis, comparison, or relative quantitation method.

The invention also provides a network, comprising a plurality of computing devices linked via a data link, such as an Ethernet cable (coax or 10BaseT), telephone line, ISDN line, wireless network, optical fiber, or other suitable signal transmission medium, whereby at least one network device (e.g., computer, disk array, etc.) comprises a pattern of magnetic domains (e.g., magnetic disk) and/or charge domains (e.g., an array of DRAM cells) composing a bit pattern encoding data acquired from an assay of the invention.

The invention also provides a method for transmitting assay data that includes generating an electronic signal on an electronic communications device, such as a modem, ISDN terminal adapter, DSL, cable modem, ATM switch, or the like, wherein the signal includes (in native or encrypted format) a bit pattern encoding data from an assay or a database comprising a plurality of assay results obtained by the method of the invention.

In a preferred embodiment, the invention provides a computer system for comparing a query target to a database containing an array of data structures, such as an assay result obtained by the method of the invention, and ranking database targets based on the degree of identity and gap weight to the target data. A central processor is preferably initialized to load and execute the computer program for alignment and/or comparison of the assay results. Data for a query target is entered into the central processor via an I/O device. Execution of the computer program results in the central processor retrieving the assay data from the data file, which comprises a binary description of an assay result.

The target data or record and the computer program can be transferred to secondary memory, which is typically random access memory (e.g., DRAM, SRAM, SGRAM, or SDRAM). Targets are ranked according to the degree of correspondence between a selected assay characteristic (e.g., binding to a selected affinity moiety) and the same characteristic of the query target and results are output via an I/O device. For example,

a central processor can be a conventional computer (e.g., Intel Pentium, PowerPC, Alpha, PA-8000, SPARC, MIPS 4400, MIPS 10000, VAX, etc.); a program can be a commercial or public domain molecular biology software package (e.g., UWGCG Sequence Analysis Software, Darwin); a data file can be an optical or magnetic disk, a data server, a memory device (e.g., DRAM, SRAM, SGRAM, SDRAM, EPROM, bubble memory, flash memory, *etc.*); an I/O device can be a terminal comprising a video display and a keyboard, a modem, an ISDN terminal adapter, an Ethernet port, a punched card reader, a magnetic strip reader, or other suitable I/O device.

The invention also preferably provides the use of a computer system, such as that described above, which comprises: (1) a computer; (2) a stored bit pattern encoding a collection of peptide sequence specificity records obtained by the methods of the invention, which may be stored in the computer; (3) a comparison target, such as a query target; and (4) a program for alignment and comparison, typically with rank-ordering of comparison results on the basis of computed similarity values.

Characteristics of breast cancer-associated proteins

Breast cancer proteins of the present invention may be classified as secreted proteins, transmembrane proteins or intracellular proteins. In one embodiment, the breast cancer protein is an intracellular protein. Intracellular proteins may be found in the cytoplasm and/or in the nucleus. Intracellular proteins are involved in all aspects of cellular function and replication (including, e.g., signaling pathways); aberrant expression of such proteins often results in unregulated or disregulated cellular processes (*see, e.g., Molecular Biology of the Cell* (Alberts, ed., 3rd ed., 1994). For example, many intracellular proteins have enzymatic activity such as protein kinase activity, protein phosphatase activity, protease activity, nucleotide cyclase activity, polymerase activity and the like. Intracellular proteins also serve as docking proteins that are involved in organizing complexes of proteins, or targeting proteins to various subcellular localizations, and are involved in maintaining the structural integrity of organelles.

An increasingly appreciated concept in characterizing proteins is the presence in the proteins of one or more motifs for which defined functions have been attributed. In addition to the highly conserved sequences found in the enzymatic domain of proteins, highly conserved sequences have been identified in proteins that are involved in protein-protein

interaction. For example, Src-homology-2 (SH2) domains bind tyrosine-phosphorylated targets in a sequence dependent manner. PTB domains, which are distinct from SH2 domains, also bind tyrosine phosphorylated targets. SH3 domains bind to proline-rich targets. In addition, PH domains, tetratricopeptide repeats and WD domains to name only a few, have been shown to mediate protein-protein interactions. Some of these may also be involved in binding to phospholipids or other second messengers. As will be appreciated by one of ordinary skill in the art, these motifs can be identified on the basis of primary sequence; thus, an analysis of the sequence of proteins may provide insight into both the enzymatic potential of the molecule and/or molecules with which the protein may associate. One useful database is Pfam (protein families), which is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains. Versions are available via the internet from Washington University in St. Louis, the Sanger Center in England, and the Karolinska Institute in Sweden (*see, e.g.*, Bateman *et al., Nuc. Acids Res.* 28:263-266 (2000); Sonnhammer *et al., Proteins* 28:405-420 (1997); Bateman *et al., Nuc. Acids Res.* 27:260-262 (1999); and Sonnhammer *et al., Nuc. Acids Res.* 26:320-322-(1998)).

In another embodiment, the breast cancer sequences are transmembrane proteins. Transmembrane proteins are molecules that span a phospholipid bilayer of a cell. They may have an intracellular domain, an extracellular domain, or both. The intracellular domains of such proteins may have a number of functions including those already described for intracellular proteins. For example, the intracellular domain may have enzymatic activity and/or may serve as a binding site for additional proteins. Frequently the intracellular domain of transmembrane proteins serves both roles. For example certain receptor tyrosine kinases have both protein kinase activity and SH2 domains. In addition, autophosphorylation of tyrosines on the receptor molecule itself, creates binding sites for additional SH2 domain containing proteins.

Transmembrane proteins may contain from one to many transmembrane domains. For example, receptor tyrosine kinases, certain cytokine receptors, receptor guanylyl cyclases and receptor serine/threonine protein kinases contain a single transmembrane domain. However, various other proteins including channels and adenylyl cyclases contain numerous transmembrane domains. Many important cell surface receptors such as G protein coupled receptors (GPCRs) are classified as "seven transmembrane

domain" proteins, as they contain 7 membrane spanning regions. Characteristics of transmembrane domains include approximately 20 consecutive hydrophobic amino acids that may be followed by charged amino acids. Therefore, upon analysis of the amino acid sequence of a particular protein, the localization and number of transmembrane domains within the protein may be predicted (*see, e.g.* PSORT web site http://psort.nibb.ac.jp/). Important transmembrane protein receptors include, but are not limited to the insulin receptor, insulin-like growth factor receptor, human growth hormone receptor, glucose transporters, transferrin receptor, epidermal growth factor receptor, low density lipoprotein receptor, epidermal growth factor receptor, leptin receptor, interleukin receptors, e.g. IL-1 receptor, IL-2 receptor,

The extracellular domains of transmembrane proteins are diverse; however, conserved motifs are found repeatedly among various extracellular domains. Conserved structure and/or functions have been ascribed to different extracellular motifs. Many extracellular domains are involved in binding to other molecules. In one aspect, extracellular domains are found on receptors. Factors that bind the receptor domain include circulating ligands, which may be peptides, proteins, or small molecules such as adenosine and the like. For example, growth factors such as EGF, FGF and PDGF are circulating growth factors that bind to their cognate receptors to initiate a variety of cellular responses. Other factors include cytokines, mitogenic factors, neurotrophic factors and the like. Extracellular domains also bind to cell-associated molecules. In this respect, they mediate cell-cell interactions. Cell-associated ligands can be tethered to the cell, e.g., via a glycosylphosphatidylinositol (GPI) anchor, or may themselves be transmembrane proteins. Extracellular domains also associate with the extracellular matrix and contribute to the maintenance of the cell structure.

Breast cancer proteins that are transmembrane are particularly preferred in the present invention as they are readily accessible targets for immunotherapeutics, as are described herein. In addition, as outlined below, transmembrane proteins can be also useful in imaging modalities. Antibodies may be used to label such readily accessible proteins *in situ*. Alternatively, antibodies can also label intracellular proteins, in which case samples are typically permeablized to provide access to intracellular proteins.

It will also be appreciated by those in the art that a transmembrane protein can be made soluble by removing transmembrane sequences, e.g., through recombinant methods.

Furthermore, transmembrane proteins that have been made soluble can be made to be secreted through recombinant means by adding an appropriate signal sequence.

In another embodiment, the breast cancer proteins are secreted proteins; the secretion of which can be either constitutive or regulated. These proteins have a signal peptide or signal sequence that targets the molecule to the secretory pathway. Secreted proteins are involved in numerous physiological events; by virtue of their circulating nature, they serve to transmit signals to various other cell types. The secreted protein may function in an autocrine manner (acting on the cell that secreted the factor), a paracrine manner (acting on cells in close proximity to the cell that secreted the factor) or an endocrine manner (acting on cells at a distance). Thus secreted molecules find use in modulating or altering numerous aspects of physiology. Breast cancer proteins that are secreted proteins are particularly preferred in the present invention as they serve as good targets for diagnostic markers, *e.g.*, for blood, plasma, serum, or stool tests.

Use of breast cancer nucleic acids

As described above, breast cancer sequence is initially identified by substantial nucleic acid and/or amino acid sequence homology or linkage to the breast cancer sequences outlined herein. Such homology can be based upon the overall nucleic acid or amino acid sequence, and is generally determined as outlined below, using either homology programs or hybridization conditions. Typically, linked sequences on a mRNA are found on the same molecule.

The breast cancer nucleic acid sequences of the invention, e.g., the sequences in Table 1-3, can be fragments of larger genes, i.e., they are nucleic acid segments. "Genes" in this context includes coding regions, non-coding regions, and mixtures of coding and non-coding regions. Accordingly, as will be appreciated by those in the art, using the sequences provided herein, extended sequences, in either direction, of the breast cancer genes can be obtained, using techniques well known in the art for cloning either longer sequences or the full length sequences; see Ausubel, *et al.*, *supra*. Much can be done by informatics and many sequences can be clustered to include multiple sequences corresponding to a single gene, e.g., systems such as UniGene (see, http://www.ncbi.nlm.nih.gov/UniGene/).

Once the breast cancer nucleic acid is identified, it can be cloned and, if necessary, its constituent parts recombined to form the entire breast cancer nucleic acid

coding regions or the entire mRNA sequence. Once isolated from its natural source, e.g., contained within a plasmid or other vector or excised therefrom as a linear nucleic acid segment, the recombinant breast cancer nucleic acid can be further-used as a probe to identify and isolate other breast cancer nucleic acids, e.g., extended coding regions. It can also be used as a "precursor" nucleic acid to make modified or variant breast cancer nucleic acids and proteins.

The breast cancer nucleic acids of the present invention are used in several ways. In a first embodiment, nucleic acid probes to the breast cancer nucleic acids are made and attached to biochips to be used in screening and diagnostic methods, as outlined below, or for administration, e.g., for gene therapy, vaccine, and/or antisense applications. Alternatively, the breast cancer nucleic acids that include coding regions of breast cancer proteins can be put into expression vectors for the expression of breast cancer proteins, again for screening purposes or for administration to a patient.

In a preferred embodiment, nucleic acid probes to breast cancer nucleic acids (both the nucleic acid sequences outlined in the figures and/or the complements thereof) are made. The nucleic acid probes attached to the biochip are designed to be substantially complementary to the breast cancer nucleic acids, *i.e.* the target sequence (either the target sequence of the sample or to other probe sequences, e.g., in sandwich assays), such that hybridization of the target sequence and the probes of the present invention occurs. As outlined below, this complementarity need not be perfect; there may be any number of base pair mismatches which will interfere with hybridization between the target sequence and the single stranded nucleic acids of the present invention. However, if the number of mutations is so great that no hybridization can occur under even the least stringent of hybridization conditions, the sequence is not a complementary target sequence. Thus, by "substantially complementary" herein is meant that the probes are sufficiently complementary to the target sequences to hybridize under normal reaction conditions, particularly high stringency conditions, as outlined herein.

A nucleic acid probe is generally single stranded but can be partially single and partially double stranded. The strandedness of the probe is dictated by the structure, composition, and properties of the target sequence. In general, the nucleic acid probes range from about 8 to about 100 bases long, with from about 10 to about 80 bases being preferred, and from about 30 to about 50 bases being particularly preferred. That is, generally whole

genes are not used. In some embodiments, much longer nucleic acids can be used, up to hundreds of bases.

In a preferred embodiment, more than one probe per sequence is used, with either overlapping probes or probes to different sections of the target being used. That is, two, three, four or more probes, with three being preferred, are used to build in a redundancy for a particular target. The probes can be overlapping (i.e., have some sequence in common), or separate. In some cases, PCR primers may be used to amplify signal for higher sensitivity.

As will be appreciated by those in the art, nucleic acids can be attached or immobilized to a solid support in a wide variety of ways. By "immobilized" and grammatical equivalents herein is meant the association or binding between the nucleic acid probe and the solid support is sufficient to be stable under the conditions of binding, washing, analysis, and removal as outlined below. The binding can typically be covalent or non-covalent. By "non-covalent binding" and grammatical equivalents herein is meant one or more of electrostatic, hydrophilic, and hydrophobic interactions. Included in non-covalent binding is the covalent attachment of a molecule, such as, streptavidin to the support and the non-covalent binding of the biotinylated probe to the streptavidin. By "covalent binding" and grammatical equivalents herein is meant that the two moieties, the solid support and the probe, are attached by at least one bond, including sigma bonds, pi bonds and coordination bonds. Covalent bonds can be formed directly between the probe and the solid support or can be formed by a cross linker or by inclusion of a specific reactive group on either the solid support or the probe or both molecules. Immobilization may also involve a combination of covalent and non-covalent interactions.

In general, the probes are attached to the biochip in a wide variety of ways, as will be appreciated by those in the art. As described herein, the nucleic acids can either be synthesized first, with subsequent attachment to the biochip, or can be directly synthesized on the biochip.

The biochip comprises a suitable solid substrate. By "substrate" or "solid support" or other grammatical equivalents herein is meant a material that can be modified to contain discrete individual sites appropriate for the attachment or association of the nucleic acid probes and is amenable to at least one detection method. As will be appreciated by those in the art, the number of possible substrates are very large, and include, but are not limited to, glass and modified or functionalized glass, plastics (including acrylics, polystyrene and

copolymers of styrene and other materials, polypropylene, polyethylene, polybutylene, polyurethanes, TeflonJ, etc.), polysaccharides, nylon or nitrocellulose, resins, silica or silica-based materials including silicon and modified silicon, carbon, metals, inorganic glasses, plastics, etc. In general, the substrates allow optical detection and do not appreciably fluoresce. A preferred substrate is described in copending application entitled Reusable Low Fluorescent Plastic Biochip, U.S. Application Serial No. 09/270,214, filed March 15, 1999, herein incorporated by reference in its entirety.

Generally the substrate is planar, although as will be appreciated by those in the art, other configurations of substrates may be used as well. For example, the probes may be placed on the inside surface of a tube, for flow-through sample analysis to minimize sample volume. Similarly, the substrate may be flexible, such as a flexible foam, including closed cell foams made of particular plastics.

In a preferred embodiment, the surface of the biochip and the probe may be derivatized with chemical functional groups for subsequent attachment of the two. Thus, e.g., the biochip is derivatized with a chemical functional group including, but not limited to, amino groups, carboxy groups, oxo groups and thiol groups, with amino groups being particularly preferred. Using these functional groups, the probes can be attached using functional groups on the probes. For example, nucleic acids containing amino groups can be attached to surfaces comprising amino groups, e.g. using linkers as are known in the art; e.g., homo-or hetero-bifunctional linkers as are well known (*see* 1994 Pierce Chemical Company catalog, technical section on cross-linkers, pages 155-200). In addition, in some cases, additional linkers, such as alkyl groups (including substituted and heteroalkyl groups) may be used.

In this embodiment, oligonucleotides are synthesized as is known in the art, and then attached to the surface of the solid support. As will be appreciated by those skilled in the art, either the 5' or 3' terminus may be attached to the solid support, or attachment may be via an internal nucleoside.

In another embodiment, the immobilization to the solid support may be very strong, yet non-covalent. For example, biotinylated oligonucleotides can be made, which bind to surfaces covalently coated with streptavidin, resulting in attachment.

Alternatively, the oligonucleotides may be synthesized on the surface, as is known in the art. For example, photoactivation techniques utilizing photopolymerization

compounds and techniques are used. In a preferred embodiment, the nucleic acids can be synthesized in situ, using well known photolithographic techniques, such as those described in WO 95/25116; WO 95/35505; U.S. Patent Nos. 5,700,637 and 5,445,934; and references cited within, all of which are expressly incorporated by reference; these methods of attachment form the basis of the Affimetrix GeneChip™ technology.

Often, amplification-based assays are performed to measure the expression level of breast cancer-associated sequences. These assays are typically performed in conjunction with reverse transcription. In such assays, a breast cancer-associated nucleic acid sequence acts as a template in an amplification reaction (e.g., Polymerase Chain Reaction, or PCR). In a quantitative amplification, the amount of amplification product will be proportional to the amount of template in the original sample. Comparison to appropriate controls provides a measure of the amount of breast cancer-associated RNA. Methods of quantitative amplification are well known to those of skill in the art. Detailed protocols for quantitative PCR are provided, e.g., in Innis *et al.*, *PCR Protocols, A Guide to Methods and Applications* (1990).

In some embodiments, a TaqMan based assay is used to measure expression. TaqMan based assays use a fluorogenic oligonucleotide probe that contains a 5' fluorescent dye and a 3' quenching agent. The probe hybridizes to a PCR product, but cannot itself be extended due to a blocking agent at the 3' end. When the PCR product is amplified in subsequent cycles, the 5' nuclease activity of the polymerase, *e.g.*, AmpliTaq, results in the cleavage of the TaqMan probe. This cleavage separates the 5' fluorescent dye and the 3' quenching agent, thereby resulting in an increase in fluorescence as a function of amplification (*see*, e.g., literature provided by Perkin-Elmer, *e.g.*, www2.perkin-elmer.com).

Other suitable amplification methods include, but are not limited to, ligase chain reaction (LCR) (*see* Wu & Wallace, *Genomics* 4:560 (1989), Landegren *et al., Science* 241:1077 (1988), and Barringer *et al.*, *Gene* 89:117 (1990)), transcription amplification (Kwoh *et al., Proc. Natl. Acad. Sci. USA* 86:1173 (1989)), self-sustained sequence replication (Guatelli *et al., Proc. Nat. Acad. Sci. USA* 87:1874 (1990)), dot PCR, and linker adapter PCR, etc.

Expression of breast cancer proteins from nucleic acids

In a preferred embodiment, breast cancer nucleic acids, e.g., encoding breast cancer proteins are used to make a variety of expression vectors to express breast cancer proteins which can then be used in screening assays, as described below. Expression vectors and recombinant DNA technology are well known to those of skill in the art (*see, e.g.,* Ausubel, *supra*, and *Gene Expression Systems* (Fernandez & Hoeffler, eds, 1999)) and are used to express proteins. The expression vectors may be either self-replicating extrachromosomal vectors or vectors which integrate into a host genome. Generally, these expression vectors include transcriptional and translational regulatory nucleic acid operably linked to the nucleic acid encoding the breast cancer protein. The term "control sequences" refers to DNA sequences used for the expression of an operably linked coding sequence in a particular host organism. Control sequences that are suitable for prokaryotes, e.g., include a promoter, optionally an operator sequence, and a ribosome binding site. Eukaryotic cells are known to utilize promoters, polyadenylation signals, and enhancers.

Nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For example, DNA for a presequence or secretory leader is operably linked to DNA for a polypeptide if it is expressed as a preprotein that participates in the secretion of the polypeptide; a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the sequence; or a ribosome binding site is operably linked to a coding sequence if it is positioned so as to facilitate translation. Generally, "operably linked" means that the DNA sequences being linked are contiguous, and, in the case of a secretory leader, contiguous and in reading phase. However, enhancers do not have to be contiguous. Linking is typically accomplished by ligation at convenient restriction sites. If such sites do not exist, synthetic oligonucleotide adaptors or linkers are used in accordance with conventional practice. Transcriptional and translational regulatory nucleic acid will generally be appropriate to the host cell used to express the breast cancer protein. Numerous types of appropriate expression vectors, and suitable regulatory sequences are known in the art for a variety of host cells.

In general, transcriptional and translational regulatory sequences may include, but are not limited to, promoter sequences, ribosomal binding sites, transcriptional start and stop sequences, translational start and stop sequences, and enhancer or activator sequences.

In a preferred embodiment, the regulatory sequences include a promoter and transcriptional start and stop sequences.

Promoter sequences encode either constitutive or inducible promoters. The promoters may be either naturally occurring promoters or hybrid promoters. Hybrid promoters, which combine elements of more than one promoter, are also known in the art, and are useful in the present invention.

In addition, an expression vector may comprise additional elements. For example, the expression vector may have two replication systems, thus allowing it to be maintained in two organisms, e.g. in mammalian or insect cells for expression and in a procaryotic host for cloning and amplification. Furthermore, for integrating expression vectors, the expression vector contains at least one sequence homologous to the host cell genome, and preferably two homologous sequences which flank the expression construct. The integrating vector may be directed to a specific locus in the host cell by selecting the appropriate homologous sequence for inclusion in the vector. Constructs for integrating vectors are well known in the art (e.g., Fernandez & Hoeffler, *supra*).

In addition, in a preferred embodiment, the expression vector contains a selectable marker gene to allow the selection of transformed host cells. Selection genes are well known in the art and will vary with the host cell used.

The breast cancer proteins of the present invention are produced by culturing a host cell transformed with an expression vector containing nucleic acid encoding a breast cancer protein, under the appropriate conditions to induce or cause expression of the breast cancer protein. Conditions appropriate for breast cancer protein expression will vary with the choice of the expression vector and the host cell, and will be easily ascertained by one skilled in the art through routine experimentation or optimization. For example, the use of constitutive promoters in the expression vector will require optimizing the growth and proliferation of the host cell, while the use of an inducible promoter requires the appropriate growth conditions for induction. In addition, in some embodiments, the timing of the harvest is important. For example, the baculoviral systems used in insect cell expression are lytic viruses, and thus harvest time selection can be crucial for product yield.

Appropriate host cells include yeast, bacteria, archaebacteria, fungi, and insect and animal cells, including mammalian cells. Of particular interest are *Saccharomyces cerevisiae* and other yeasts, *E. coli, Bacillus subtilis*, Sf9 cells, C129 cells, 293 cells,

Neurospora, BHK, CHO, COS, HeLa cells, HUVEC (human umbilical vein endothelial cells), THP1 cells (a macrophage cell line) and various other human cells and cell lines.

In a preferred embodiment, the breast cancer proteins are expressed in mammalian cells. Mammalian expression systems are also known in the art, and include retroviral and adenoviral systems. One expression vector system is a retroviral vector system such as is generally described in PCT/US97/01019 and PCT/US97/01048, both of which are hereby expressly incorporated by reference. Of particular use as mammalian promoters are the promoters from mammalian viral genes, since the viral genes are often highly expressed and have a broad host range. Examples include the SV40 early promoter, mouse mammary tumor virus LTR promoter, adenovirus major late promoter, herpes simplex virus promoter, and the CMV promoter (*see, e.g.,* Fernandez & Hoeffler, *supra*). Typically, transcription termination and polyadenylation sequences recognized by mammalian cells are regulatory regions located 3' to the translation stop codon and thus, together with the promoter elements, flank the coding sequence. Examples of transcription terminator and polyadenlyation signals include those derived form SV40.

The methods of introducing exogenous nucleic acid into mammalian hosts, as well as other hosts, is well known in the art, and will vary with the host cell used. Techniques include dextran-mediated transfection, calcium phosphate precipitation, polybrene mediated transfection, protoplast fusion, electroporation, viral infection, encapsulation of the polynucleotide(s) in liposomes, and direct microinjection of the DNA into nuclei.

In a preferred embodiment, breast cancer proteins are expressed in bacterial systems. Bacterial expression systems are well known in the art. Promoters from bacteriophage may also be used and are known in the art. In addition, synthetic promoters and hybrid promoters are also useful; e.g., the tac promoter is a hybrid of the trp and lac promoter sequences. Furthermore, a bacterial promoter can include naturally occurring promoters of non-bacterial origin that have the ability to bind bacterial RNA polymerase and initiate transcription. In addition to a functioning promoter sequence, an efficient ribosome binding site is desirable. The expression vector may also include a signal peptide sequence that provides for secretion of the breast cancer protein in bacteria. The protein is either secreted into the growth media (gram-positive bacteria) or into the periplasmic space, located between the inner and outer membrane of the cell (gram-negative bacteria). The bacterial

expression vector may also include a selectable marker gene to allow for the selection of bacterial strains that have been transformed. Suitable selection genes include genes which render the bacteria resistant to drugs such as ampicillin, chloramphenicol, erythromycin, kanamycin, neomycin and tetracycline. Selectable markers also include biosynthetic genes, such as those in the histidine, tryptophan and leucine biosynthetic pathways. These components are assembled into expression vectors. Expression vectors for bacteria are well known in the art, and include vectors for *Bacillus subtilis, E. coli, Streptococcus cremoris*, and S*treptococcus lividans*, among others (e.g., Fernandez & Hoeffler, *supra*). The bacterial expression vectors are transformed into bacterial host cells using techniques well known in the art, such as calcium chloride treatment, electroporation, and others.

In one embodiment, breast cancer proteins are produced in insect cells. Expression vectors for the transformation of insect cells, and in particular, baculovirus-based expression vectors, are well known in the art.

In a preferred embodiment, breast cancer protein is produced in yeast cells. Yeast expression systems are well known in the art, and include expression vectors for *Saccharomyces cerevisiae, Candida albicans* and *C. maltosa, Hansenula polymorpha, Kluyveromyces fragilis* and *K. lactis, Pichia guillerimondii and P. pastoris, Schizosaccharomyces pombe*, and *Yarrowia lipolytica*.

The breast cancer protein may also be made as a fusion protein, using techniques well known in the art. Thus, e.g., for the creation of monoclonal antibodies, if the desired epitope is small, the breast cancer protein may be fused to a carrier protein to form an immunogen. Alternatively, the breast cancer protein may be made as a fusion protein to increase expression, or for other reasons. For example, when the breast cancer protein is a breast cancer peptide, the nucleic acid encoding the peptide may be linked to other nucleic acid for expression purposes.

In a preferred embodiment, the breast cancer protein is purified or isolated after expression. Breast cancer proteins may be isolated or purified in a variety of ways known to those skilled in the art depending on what other components are present in the sample. Standard purification methods include electrophoretic, molecular, immunological and chromatographic techniques, including ion exchange, hydrophobic, affinity, and reverse-phase HPLC chromatography, and chromatofocusing. For example, the breast cancer protein may be purified using a standard anti-breast cancer protein antibody column. Ultrafiltration

and diafiltration techniques, in conjunction with protein concentration, are also useful. For general guidance in suitable purification techniques, see Scopes, *Protein Purification* (1982). The degree of purification necessary will vary depending on the use of the breast cancer protein. In some instances no purification will be necessary.

Once expressed and purified if necessary, the breast cancer proteins and nucleic acids are useful in a number of applications. They may be used as immunoselection reagents, as vaccine reagents, as screening agents, etc.

Variants of breast cancer proteins

In one embodiment, the breast cancer proteins are derivative or variant breast cancer proteins as compared to the wild-type sequence. That is, as outlined more fully below, the derivative breast cancer peptide will often contain at least one amino acid substitution, deletion or insertion, with amino acid substitutions being particularly preferred. The amino acid substitution, insertion or deletion may occur at any residue within the breast cancer peptide.

Also included within one embodiment of breast cancer proteins of the present invention are amino acid sequence variants. These variants typically fall into one or more of three classes: substitutional, insertional or deletional variants. These variants ordinarily are prepared by site specific mutagenesis of nucleotides in the DNA encoding the breast cancer protein, using cassette or PCR mutagenesis or other techniques well known in the art, to produce DNA encoding the variant, and thereafter expressing the DNA in recombinant cell culture as outlined above. However, variant breast cancer protein fragments having up to about 100-150 residues may be prepared by in vitro synthesis using established techniques. Amino acid sequence variants are characterized by the predetermined nature of the variation, a feature that sets them apart from naturally occurring allelic or interspecies variation of the breast cancer protein amino acid sequence. The variants typically exhibit the same qualitative biological activity as the naturally occurring analogue, although variants can also be selected which have modified characteristics as will be more fully outlined below.

While the site or region for introducing an amino acid sequence variation is predetermined, the mutation per se need not be predetermined. For example, in order to optimize the performance of a mutation at a given site, random mutagenesis may be conducted at the target codon or region and the expressed breast cancer variants screened for

the optimal combination of desired activity. Techniques for making substitution mutations at predetermined sites in DNA having a known sequence are well known, e.g., M13 primer mutagenesis and PCR mutagenesis. Screening of the mutants is done using assays of breast cancer protein activities.

Amino acid substitutions are typically of single residues; insertions usually will be on the order of from about 1 to 20 amino acids, although considerably larger insertions may be tolerated. Deletions range from about 1 to about 20 residues, although in some cases deletions may be much larger.

Substitutions, deletions, insertions or any combination thereof may be used to arrive at a final derivative. Generally these changes are done on a few amino acids to minimize the alteration of the molecule. However, larger changes may be tolerated in certain circumstances. When small alterations in the characteristics of the breast cancer protein are desired, substitutions are generally made in accordance with the amino acid substitution relationships provided in the definition section.

The variants typically exhibit the same qualitative biological activity and will elicit the same immune response as the naturally-occurring analog, although variants also are selected to modify the characteristics of the breast cancer proteins as needed. Alternatively, the variant may be designed such that the biological activity of the breast cancer protein is altered. For example, glycosylation sites may be altered or removed.

Substantial changes in function or immunological identity are made by selecting substitutions that are less conservative than those described above. For example, substitutions may be made which more significantly affect: the structure of the polypeptide backbone in the area of the alteration, for example the alpha-helical or beta-sheet structure; the charge or hydrophobicity of the molecule at the target site; or the bulk of the side chain. The substitutions which in general are expected to produce the greatest changes in the polypeptide's properties are those in which (a) a hydrophilic residue, e.g. seryl or threonyl is substituted for (or by) a hydrophobic residue, e.g. leucyl, isoleucyl, phenylalanyl, valyl or alanyl; (b) a cysteine or proline is substituted for (or by) any other residue; (c) a residue having an electropositive side chain, e.g. lysyl, arginyl, or histidyl, is substituted for (or by) an electronegative residue, e.g. glutamyl or aspartyl; or (d) a residue having a bulky side chain, e.g. phenylalanine, is substituted for (or by) one not having a side chain, e.g. glycine.

Covalent modifications of breast cancer polypeptides are included within the scope of this invention. One type of covalent modification includes reacting targeted amino acid residues of a breast cancer polypeptide with an organic derivatizing agent that is capable of reacting with selected side chains or the N-or C-terminal residues of a breast cancer polypeptide. Derivatization with bifunctional agents is useful, for instance, for crosslinking breast cancer polypeptides to a water-insoluble support matrix or surface for use in the method for purifying anti-breast cancer polypeptide antibodies or screening assays, as is more fully described below. Commonly used crosslinking agents include, e.g., 1,1-bis(diazoacetyl)-2-phenylethane, glutaraldehyde, N-hydroxysuccinimide esters, e.g., esters with 4-azidosalicylic acid, homobifunctional imidoesters, including disuccinimidyl esters such as 3,3'-dithiobis(succinimidylpropionate), bifunctional maleimides such as bis-N-maleimido-1,8-octane and agents such as methyl-3-((p-azidophenyl)dithio)propioimidate.

Other modifications include deamidation of glutaminyl and asparaginyl residues to the corresponding glutamyl and aspartyl residues, respectively, hydroxylation of proline and lysine, phosphorylation of hydroxyl groups of seryl, threonyl or tyrosyl residues, methylation of the amino groups of the lysine, arginine, and histidine side chains (Creighton, *Proteins: Structure and Molecular Properties*, pp. 79-86 (1983)), acetylation of the N-terminal amine, and amidation of any C-terminal carboxyl group.

Another type of covalent modification of the breast cancer polypeptide included within the scope of this invention comprises altering the native glycosylation pattern of the polypeptide. "Altering the native glycosylation pattern" is intended for purposes herein to mean deleting one or more carbohydrate moieties found in native sequence breast cancer polypeptide, and/or adding one or more glycosylation sites that are not present in the native sequence breast cancer polypeptide. Glycosylation patterns can be altered in many ways. For example the use of different cell types to express breast cancer-associated sequences can result in different glycosylation patterns.

Addition of glycosylation sites to breast cancer polypeptides may also be accomplished by altering the amino acid sequence thereof. The alteration may be made, e.g., by the addition of, or substitution by, one or more serine or threonine residues to the native sequence breast cancer polypeptide (for O-linked glycosylation sites). The breast cancer amino acid sequence may optionally be altered through changes at the DNA level,

particularly by mutating the DNA encoding the breast cancer polypeptide at preselected bases such that codons are generated that will translate into the desired amino acids.

Another means of increasing the number of carbohydrate moieties on the breast cancer polypeptide is by chemical or enzymatic coupling of glycosides to the polypeptide. Such methods are described in the art, e.g., in WO 87/05330, and in Aplin & Wriston, *CRC Crit. Rev. Biochem.*, pp. 259-306 (1981).

Removal of carbohydrate moieties present on the breast cancer polypeptide may be accomplished chemically or enzymatically or by mutational substitution of codons encoding for amino acid residues that serve as targets for glycosylation. Chemical deglycosylation techniques are known in the art and described, for instance, by Hakimuddin, *et al.*, *Arch. Biochem. Biophys.*, 259:52 (1987) and by Edge *et al.*, *Anal. Biochem.*, 118:131 (1981). Enzymatic cleavage of carbohydrate moieties on polypeptides can be achieved by the use of a variety of endo-and exo-glycosidases as described by Thotakura *et al.*, *Meth. Enzymol.*, 138:350 (1987).

Another type of covalent modification of breast cancer comprises linking the breast cancer polypeptide to one of a variety of nonproteinaceous polymers, e.g., polyethylene glycol, polypropylene glycol, or polyoxyalkylenes, in the manner set forth in U.S. Patent Nos. 4,640,835; 4,496,689; 4,301,144; 4,670,417; 4,791,192 or 4,179,337.

Breast cancer polypeptides of the present invention may also be modified in a way to form chimeric molecules comprising a breast cancer polypeptide fused to another, heterologous polypeptide or amino acid sequence. In one embodiment, such a chimeric molecule comprises a fusion of a breast cancer polypeptide with a tag polypeptide which provides an epitope to which an anti-tag antibody can selectively bind. The epitope tag is generally placed at the amino-or carboxyl-terminus of the breast cancer polypeptide. The presence of such epitope-tagged forms of a breast cancer polypeptide can be detected using an antibody against the tag polypeptide. Also, provision of the epitope tag enables the breast cancer polypeptide to be readily purified by affinity purification using an anti-tag antibody or another type of affinity matrix that binds to the epitope tag. In an alternative embodiment, the chimeric molecule may comprise a fusion of a breast cancer polypeptide with an immunoglobulin or a particular region of an immunoglobulin. For a bivalent form of the chimeric molecule, such a fusion could be to the Fc region of an IgG molecule.

Various tag polypeptides and their respective antibodies are well known in the art. Examples include poly-histidine (poly-his) or poly-histidine-glycine (poly-his-gly) tags; HIS6 and metal chelation tags, the flu HA tag polypeptide and its antibody 12CA5 (Field *et al., Mol. Cell. Biol.* 8:2159-2165 (1988)); the c-myc tag and the 8F9, 3C7, 6E10, G4, B7 and 9E10 antibodies thereto (Evan *et al., Molecular and Cellular Biology* 5:3610-3616 (1985)); and the Herpes Simplex virus glycoprotein D (gD) tag and its antibody (*Paborsky et al., Protein Engineering* 3(6):547-553 (1990)). Other tag polypeptides include the Flag-peptide (*Hopp et al., BioTechnology* 6:1204-1210 (1988)); the KT3 epitope peptide (Martin *et al., Science* 255:192-194 (1992)); tubulin epitope peptide (Skinner *et al., J. Biol. Chem.* 266:15163-15166 (1991)); and the T7 gene 10 protein peptide tag (Lutz-Freyermuth *et al., Proc. Natl. Acad. Sci. USA* 87:6393-6397 (1990)).

Also included are other breast cancer proteins of the breast cancer family, and breast cancer proteins from other organisms, which are cloned and expressed as outlined below. Thus, probe or degenerate polymerase chain reaction (PCR) primer sequences may be used to find other related breast cancer proteins from humans or other organisms. As will be appreciated by those in the art, particularly useful probe and/or PCR primer sequences include the unique areas of the breast cancer nucleic acid sequence. As is generally known in the art, preferred PCR primers are from about 15 to about 35 nucleotides in length, with from about 20 to about 30 being preferred, and may contain inosine as needed. The conditions for the PCR reaction are well known in the art (e.g., Innis, PCR Protocols, *supra*).

Antibodies to breast cancer proteins

In a preferred embodiment, when the breast cancer protein is to be used to generate antibodies, e.g., for immunotherapy or immunodiagnosis, the breast cancer protein should share at least one epitope or determinant with the full length protein. By "epitope" or "determinant" herein is typically meant a portion of a protein which will generate and/or bind an antibody or T-cell receptor in the context of MHC. Thus, in most instances, antibodies made to a smaller breast cancer protein will be able to bind to the full-length protein, particularly linear epitopes. In a preferred embodiment, the epitope is unique; that is, antibodies generated to a unique epitope show little or no cross-reactivity.

Methods of preparing polyclonal antibodies are known to the skilled artisan (e.g., Coligan, *supra*; and Harlow & Lane, *supra*). Polyclonal antibodies can be raised in a

mammal, e.g., by one or more injections of an immunizing agent and, if desired, an adjuvant. Typically, the immunizing agent and/or adjuvant will be injected in the mammal by multiple subcutaneous or intraperitoneal injections. The immunizing agent may include a protein encoded by a nucleic acid of the figures or fragment thereof or a fusion protein thereof. It may be useful to conjugate the immunizing agent to a protein known to be immunogenic in the mammal being immunized. Examples of such immunogenic proteins include but are not limited to keyhole limpet hemocyanin, serum albumin, bovine thyroglobulin, and soybean trypsin inhibitor. Examples of adjuvants which may be employed include Freund's complete adjuvant and MPL-TDM adjuvant (monophosphoryl Lipid A, synthetic trehalose dicorynomycolate). The immunization protocol may be selected by one skilled in the art without undue experimentation.

The antibodies may, alternatively, be monoclonal antibodies. Monoclonal antibodies may be prepared using hybridoma methods, such as those described by Kohler & Milstein, *Nature* 256:495 (1975). In a hybridoma method, a mouse, hamster, or other appropriate host animal, is typically immunized with an immunizing agent to elicit lymphocytes that produce or are capable of producing antibodies that will specifically bind to the immunizing agent. Alternatively, the lymphocytes may be immunized in vitro. The immunizing agent will typically include a polypeptide encoded by a nucleic acid of Tables 1 3 or fragment thereof, or a fusion protein thereof. Generally, either peripheral blood lymphocytes ("PBLs") are used if cells of human origin are desired, or spleen cells or lymph node cells are used if non-human mammalian sources are desired. The lymphocytes are then fused with an immortalized cell line using a suitable fusing agent, such as polyethylene glycol, to form a hybridoma cell (Goding, *Monoclonal Antibodies: Principles and Practice*, pp. 59-103 (1986)). Immortalized cell lines are usually transformed mammalian cells, particularly myeloma cells of rodent, bovine and human origin. Usually, rat or mouse myeloma cell lines are employed. The hybridoma cells may be cultured in a suitable culture medium that preferably contains one or more substances that inhibit the growth or survival of the unfused, immortalized cells. For example, if the parental cells lack the enzyme hypoxanthine guanine phosphoribosyl transferase (HGPRT or HPRT), the culture medium for the hybridomas typically will include hypoxanthine, aminopterin, and thymidine ("HAT medium"), which substances prevent the growth of HGPRT-deficient cells.

In one embodiment, the antibodies are bispecific antibodies. Bispecific antibodies are monoclonal, preferably human or humanized, antibodies that have binding specificities for at least two different antigens or that have binding specificities for two epitopes on the same antigen. In one embodiment, one of the binding specificities is for a protein encoded by a nucleic acid Table 1-3 or a fragment thereof, the other one is for any other antigen, and preferably for a cell-surface protein or receptor or receptor subunit, preferably one that is tumor specific. Alternatively, tetramer-type technology may create multivalent reagents.

In a preferred embodiment, the antibodies to breast cancer protein are capable of reducing or eliminating a biological function of a breast cancer protein, as is described below. That is, the addition of anti-breast cancer protein antibodies (either polyclonal or preferably monoclonal) to breast cancer tissue (or cells containing breast cancer) may reduce or eliminate the breast cancer. Generally, at least a 25% decrease in activity, growth, size or the like is preferred, with at least about 50% being particularly preferred and about a 95-100% decrease being especially preferred.

In a preferred embodiment the antibodies to the breast cancer proteins are humanized antibodies (e.g., Xenerex Biosciences, Mederex, Inc., Abgenix, Inc., Protein Design Labs,Inc.) Humanized forms of non-human (e.g., murine) antibodies are chimeric molecules of immunoglobulins, immunoglobulin chains or fragments thereof (such as Fv, Fab, Fab', F(ab')2 or other antigen-binding subsequences of antibodies) which contain minimal sequence derived from non-human immunoglobulin. Humanized antibodies include human immunoglobulins (recipient antibody) in which residues from a complementary determining region (CDR) of the recipient are replaced by residues from a CDR of a non-human species (donor antibody) such as mouse, rat or rabbit having the desired specificity, affinity and capacity. In some instances, Fv framework residues of the human immunoglobulin are replaced by corresponding non-human residues. Humanized antibodies may also comprise residues which are found neither in the recipient antibody nor in the imported CDR or framework sequences. In general, a humanized antibody will comprise substantially all of at least one, and typically two, variable domains, in which all or substantially all of the CDR regions correspond to those of a non-human immunoglobulin and all or substantially all of the framework (FR) regions are those of a human immunoglobulin consensus sequence. The humanized antibody optimally also will comprise

at least a portion of an immunoglobulin constant region (Fc), typically that of a human immunoglobulin (Jones *et al.*, *Nature* 321:522-525 (1986); Riechmann *et al.*, *Nature* 332:323-329 (1988); and Presta, *Curr. Op. Struct. Biol.* 2:593-596 (1992)). Humanization can be essentially performed following the method of Winter and co-workers (Jones *et al.*, *Nature* 321:522-525 (1986); Riechmann *et al.*, *Nature* 332:323-327 (1988); Verhoeyen *et al.*, *Science* 239:1534-1536 (1988)), by substituting rodent CDRs or CDR sequences for the corresponding sequences of a human antibody. Accordingly, such humanized antibodies are chimeric antibodies (U.S. Patent No. 4,816,567), wherein substantially less than an intact human variable domain has been substituted by the corresponding sequence from a non-human species.

Human antibodies can also be produced using various techniques known in the art, including phage display libraries (Hoogenboom & Winter, *J. Mol. Biol.* 227:381 (1991); Marks *et al.*, *J. Mol. Biol.* 222:581 (1991)). The techniques of Cole *et al.* and Boerner *et al.* are also available for the preparation of human monoclonal antibodies (Cole *et al.*, *Monoclonal Antibodies and Cancer Therapy,* p. 77 (1985) and Boerner *et al.*, *J. Immunol.* 147(1):86-95 (1991)). Similarly, human antibodies can be made by introducing of human immunoglobulin loci into transgenic animals, e.g., mice in which the endogenous immunoglobulin genes have been partially or completely inactivated. Upon challenge, human antibody production is observed, which closely resembles that seen in humans in all respects, including gene rearrangement, assembly, and antibody repertoire. This approach is described, e.g., in U.S. Patent Nos. 5,545,807; 5,545,806; 5,569,825; 5,625,126; 5,633,425; 5,661,016, and in the following scientific publications: Marks *et al.*, *Bio/Technology* 10:779-783 (1992); Lonberg *et al.*, *Nature* 368:856-859 (1994); Morrison, *Nature* 368:812-13 (1994); Fishwild *et al.*, *Nature Biotechnology* 14:845-51 (1996); Neuberger, *Nature Biotechnology* 14:826 (1996); Lonberg & Huszar, *Intern. Rev. Immunol.* 13:65-93 (1995).

By immunotherapy is meant treatment of breast cancer with an antibody raised against breast cancer proteins. As used herein, immunotherapy can be passive or active. Passive immunotherapy as defined herein is the passive transfer of antibody to a recipient (patient). Active immunization is the induction of antibody and/or T-cell responses in a recipient (patient). Induction of an immune response is the result of providing the recipient with an antigen to which antibodies are raised. As appreciated by one of ordinary skill in the art, the antigen may be provided by injecting a polypeptide against which antibodies are

desired to be raised into a recipient, or contacting the recipient with a nucleic acid capable of expressing the antigen and under conditions for expression of the antigen, leading to an immune response.

In a preferred embodiment the breast cancer proteins against which antibodies are raised are secreted proteins as described above. Without being bound by theory, antibodies used for treatment, bind and prevent the secreted protein from binding to its receptor, thereby inactivating the secreted breast cancer protein.

In another preferred embodiment, the breast cancer protein to which antibodies are raised is a transmembrane protein. Without being bound by theory, antibodies used for treatment, bind the extracellular domain of the breast cancer protein and prevent it from binding to other proteins, such as circulating ligands or cell-associated molecules. The antibody may cause down-regulation of the transmembrane breast cancer protein. As will be appreciated by one of ordinary skill in the art, the antibody may be a competitive, non-competitive or uncompetitive inhibitor of protein binding to the extracellular domain of the breast cancer protein. The antibody is also an antagonist of the breast cancer protein. Further, the antibody prevents activation of the transmembrane breast cancer protein. In one aspect, when the antibody prevents the binding of other molecules to the breast cancer protein, the antibody prevents growth of the cell. The antibody may also be used to target or sensitize the cell to cytotoxic agents, including, but not limited to TNF-α, TNF-β, IL-1, INF-γ and IL-2, or chemotherapeutic agents including 5FU, vinblastine, actinomycin D, cisplatin, methotrexate, and the like. In some instances the antibody belongs to a sub-type that activates serum complement when complexed with the transmembrane protein thereby mediating cytotoxicity or antigen-dependent cytotoxicity (ADCC). Thus, breast cancer is treated by administering to a patient antibodies directed against the transmembrane breast cancer protein. Antibody-labeling may activate a co-toxin, localize a toxin payload, or otherwise provide means to locally ablate cells.

In another preferred embodiment, the antibody is conjugated to an effector moiety. The effector moiety can be any number of molecules, including labelling moieties such as radioactive labels or fluorescent labels, or can be a therapeutic moiety. In one aspect the therapeutic moiety is a small molecule that modulates the activity of the breast cancer protein. In another aspect the therapeutic moiety modulates the activity of molecules associated with or in close proximity to the breast cancer protein. The therapeutic moiety

may inhibit enzymatic activity such as protease or collagenase or protein kinase activity associated with breast cancer.

In a preferred embodiment, the therapeutic moiety can also be a cytotoxic agent. In this method, targeting the cytotoxic agent to breast cancer tissue or cells, results in a reduction in the number of afflicted cells, thereby reducing symptoms associated with breast cancer. Cytotoxic agents are numerous and varied and include, but are not limited to, cytotoxic drugs or toxins or active fragments of such toxins. Suitable toxins and their corresponding fragments include diphtheria A chain, exotoxin A chain, ricin A chain, abrin A chain, curcin, crotin, phenomycin, enomycin and the like. Cytotoxic agents also include radiochemicals made by conjugating radioisotopes to antibodies raised against breast cancer proteins, or binding of a radionuclide to a chelating agent that has been covalently attached to the antibody. Targeting the therapeutic moiety to transmembrane breast cancer proteins not only serves to increase the local concentration of therapeutic moiety in the breast cancer afflicted area, but also serves to reduce deleterious side effects that may be associated with the therapeutic moiety.

In another preferred embodiment, the breast cancer protein against which the antibodies are raised is an intracellular protein. In this case, the antibody may be conjugated to a protein which facilitates entry into the cell. In one case, the antibody enters the cell by endocytosis. In another embodiment, a nucleic acid encoding the antibody is administered to the individual or cell. Moreover, wherein the breast cancer protein can be targeted within a cell, i.e., the nucleus, an antibody thereto contains a signal for that target localization, i.e., a nuclear localization signal.

The breast cancer antibodies of the invention specifically bind to breast cancer proteins. By "specifically bind" herein is meant that the antibodies bind to the protein with a K_d of at least about 0.1 mM, more usually at least about 1 μM, preferably at least about 0.1 μM or better, and most preferably, 0.01 μM or better. Selectivity of binding is also important.

Detection of breast cancer sequence for diagnostic and therapeutic applications

In one aspect, the RNA expression levels of genes are determined for different cellular states in the breast cancer phenotype. Expression levels of genes in normal tissue (i.e., not undergoing breast cancer) and in breast cancer tissue (and in some cases, for varying

severities of breast cancer that relate to prognosis, as outlined below) are evaluated to provide expression profiles. An expression profile of a particular cell state or point of development is essentially a "fingerprint" of the state. While two states may have any particular gene similarly expressed, the evaluation of a number of genes simultaneously allows the generation of a gene expression profile that is reflective of the state of the cell. By comparing expression profiles of cells in different states, information regarding which genes are important (including both up- and down-regulation of genes) in each of these states is obtained. Then, diagnosis may be performed or confirmed to determine whether a tissue sample has the gene expression profile of normal or cancerous tissue. This will provide for molecular diagnosis of related conditions.

"Differential expression," or grammatical equivalents as used herein, refers to qualitative or quantitative differences in the temporal and/or cellular gene expression patterns within and among cells and tissue. Thus, a differentially expressed gene can qualitatively have its expression altered, including an activation or inactivation, in, e.g., normal versus breast cancer tissue. Genes may be turned on or turned off in a particular state, relative to another state thus permitting comparison of two or more states. A qualitatively regulated gene will exhibit an expression pattern within a state or cell type which is detectable by standard techniques. Some genes will be expressed in one state or cell type, but not in both. Alternatively, the difference in expression may be quantitative, e.g., in that expression is increased or decreased; i.e., gene expression is either upregulated, resulting in an increased amount of transcript, or downregulated, resulting in a decreased amount of transcript. The degree to which expression differs need only be large enough to quantify via standard characterization techniques as outlined below, such as by use of Affymetrix GeneChip™ expression arrays, Lockhart, *Nature Biotechnology* 14:1675-1680 (1996), hereby expressly incorporated by reference. Other techniques include, but are not limited to, quantitative reverse transcriptase PCR, northern analysis and RNase protection. As outlined above, preferably the change in expression (i.e., upregulation or downregulation) is at least about 50%, more preferably at least about 100%, more preferably at least about 150%, more preferably at least about 200%, with from 300 to at least 1000% being especially preferred.

Evaluation may be at the gene transcript, or the protein level. The amount of gene expression may be monitored using nucleic acid probes to the DNA or RNA equivalent of the gene transcript, and the quantification of gene expression levels, or, alternatively, the

final gene product itself (protein) can be monitored, e.g., with antibodies to the breast cancer protein and standard immunoassays (ELISAs, etc.) or other techniques, including mass spectroscopy assays, 2D gel electrophoresis assays, etc. Proteins corresponding to breast cancer genes, i.e., those identified as being important in a breast cancer phenotype, can be evaluated in a breast cancer diagnostic test.

In a preferred embodiment, gene expression monitoring is performed simultaneously on a number of genes. Multiple protein expression monitoring can be performed as well. Similarly, these assays may be performed on an individual basis as well.

In this embodiment, the breast cancer nucleic acid probes are attached to biochips as outlined herein for the detection and quantification of breast cancer sequences in a particular cell. The assays are further described below in the example. PCR techniques can be used to provide greater sensitivity.

In a preferred embodiment nucleic acids encoding the breast cancer protein are detected. Although DNA or RNA encoding the breast cancer protein may be detected, of particular interest are methods wherein an mRNA encoding a breast cancer protein is detected. Probes to detect mRNA can be a nucleotide/deoxynucleotide probe that is complementary to and hybridizes with the mRNA and includes, but is not limited to, oligonucleotides, cDNA or RNA. Probes also should contain a detectable label, as defined herein. In one method the mRNA is detected after immobilizing the nucleic acid to be examined on a solid support such as nylon membranes and hybridizing the probe with the sample. Following washing to remove the non-specifically bound probe, the label is detected. In another method detection of the mRNA is performed in situ. In this method permeabilized cells or tissue samples are contacted with a detectably labeled nucleic acid probe for sufficient time to allow the probe to hybridize with the target mRNA. Following washing to remove the non-specifically bound probe, the label is detected. For example a digoxygenin labeled riboprobe (RNA probe) that is complementary to the mRNA encoding a breast cancer protein is detected by binding the digoxygenin with an anti-digoxygenin secondary antibody and developed with nitro blue tetrazolium and 5-bromo-4-chloro-3-indoyl phosphate.

In a preferred embodiment, various proteins from the three classes of proteins as described herein (secreted, transmembrane or intracellular proteins) are used in diagnostic assays. The breast cancer proteins, antibodies, nucleic acids, modified proteins and cells

containing breast cancer sequences are used in diagnostic assays. This can be performed on an individual gene or corresponding polypeptide level. In a preferred embodiment, the expression profiles are used, preferably in conjunction with high throughput screening techniques to allow monitoring for expression profile genes and/or corresponding polypeptides.

As described and defined herein, breast cancer proteins, including intracellular, transmembrane or secreted proteins, find use as markers of breast cancer. Detection of these proteins in putative breast cancer tissue allows for detection or diagnosis of breast cancer. In one embodiment, antibodies are used to detect breast cancer proteins. A preferred method separates proteins from a sample by electrophoresis on a gel (typically a denaturing and reducing protein gel, but may be another type of gel, including isoelectric focusing gels and the like). Following separation of proteins, the breast cancer protein is detected, e.g., by immunoblotting with antibodies raised against the breast cancer protein. Methods of immunoblotting are well known to those of ordinary skill in the art.

In another preferred method, antibodies to the breast cancer protein find use in *in situ* imaging techniques, e.g., in histology (e.g., *Methods in Cell Biology: Antibodies in Cell Biology*, volume 37 (Asai, ed. 1993)). In this method cells are contacted with from one to many antibodies to the breast cancer protein(s). Following washing to remove non-specific antibody binding, the presence of the antibody or antibodies is detected. In one embodiment the antibody is detected by incubating with a secondary antibody that contains a detectable label. In another method the primary antibody to the breast cancer protein(s) contains a detectable label, e.g. an enzyme marker that can act on a substrate. In another preferred embodiment each one of multiple primary antibodies contains a distinct and detectable label. This method finds particular use in simultaneous screening for a plurality of breast cancer proteins. As will be appreciated by one of ordinary skill in the art, many other histological imaging techniques are also provided by the invention.

In a preferred embodiment the label is detected in a fluorometer which has the ability to detect and distinguish emissions of different wavelengths. In addition, a fluorescence activated cell sorter (FACS) can be used in the method.

In another preferred embodiment, antibodies find use in diagnosing breast cancer from blood, serum, plasma, stool, and other samples. Such samples, therefore, are useful as samples to be probed or tested for the presence of breast cancer proteins.

Antibodies can be used to detect a breast cancer protein by previously described immunoassay techniques including ELISA, immunoblotting (western blotting), immunoprecipitation, BIACORE technology and the like. Conversely, the presence of antibodies may indicate an immune response against an endogenous breast cancer protein.

In a preferred embodiment, *in situ* hybridization of labeled breast cancer nucleic acid probes to tissue arrays is done. For example, arrays of tissue samples, including breast cancer tissue and/or normal tissue, are made. *In situ* hybridization (*see, e.g.,* Ausubel, *supra*) is then performed. When comparing the fingerprints between an individual and a standard, the skilled artisan can make a diagnosis, a prognosis, or a prediction based on the findings. It is further understood that the genes which indicate the diagnosis may differ from those which indicate the prognosis and molecular profiling of the condition of the cells may lead to distinctions between responsive or refractory conditions or may be predictive of outcomes.

In a preferred embodiment, the breast cancer proteins, antibodies, nucleic acids, modified proteins and cells containing breast cancer sequences are used in prognosis assays. As above, gene expression profiles can be generated that correlate to breast cancer, in terms of long term prognosis. Again, this may be done on either a protein or gene level, with the use of genes being preferred. As above, breast cancer probes may be attached to biochips for the detection and quantification of breast cancer sequences in a tissue or patient. The assays proceed as outlined above for diagnosis. PCR method may provide more sensitive and accurate quantification.

Assays for therapeutic compounds

In a preferred embodiment members of the proteins, nucleic acids, and antibodies as described herein are used in drug screening assays. The breast cancer proteins, antibodies, nucleic acids, modified proteins and cells containing breast cancer sequences are used in drug screening assays or by evaluating the effect of drug candidates on a "gene expression profile" or expression profile of polypeptides. In a preferred embodiment, the expression profiles are used, preferably in conjunction with high throughput screening techniques to allow monitoring for expression profile genes after treatment with a candidate agent (e.g., Zlokarnik, *et al.*, *Science* 279:84-8 (1998); Heid, *Genome Res* 6:986-94, 1996).

In a preferred embodiment, the breast cancer proteins, antibodies, nucleic acids, modified proteins and cells containing the native or modified breast cancer proteins are used in screening assays. That is, the present invention provides novel methods for screening for compositions which modulate the breast cancer phenotype or an identified physiological function of a breast cancer protein. As above, this can be done on an individual gene level or by evaluating the effect of drug candidates on a "gene expression profile". In a preferred embodiment, the expression profiles are used, preferably in conjunction with high throughput screening techniques to allow monitoring for expression profile genes after treatment with a candidate agent, see Zlokarnik, *supra*.

Having identified the differentially expressed genes herein, a variety of assays may be executed. In a preferred embodiment, assays may be run on an individual gene or protein level. That is, having identified a particular gene as up regulated in breast cancer, test compounds can be screened for the ability to modulate gene expression or for binding to the breast cancer protein. "Modulation" thus includes both an increase and a decrease in gene expression. The preferred amount of modulation will depend on the original change of the gene expression in normal versus tissue undergoing breast cancer, with changes of at least 10%, preferably 50%, more preferably 100-300%, and in some embodiments 300-1000% or greater. Thus, if a gene exhibits a 4-fold increase in breast cancer tissue compared to normal tissue, a decrease of about four-fold is often desired; similarly, a 10-fold decrease in breast cancer tissue compared to normal tissue often provides a target value of a 10-fold increase in expression to be induced by the test compound.

The amount of gene expression may be monitored using nucleic acid probes and the quantification of gene expression levels, or, alternatively, the gene product itself can be monitored, e.g., through the use of antibodies to the breast cancer protein and standard immunoassays. Proteomics and separation techniques may also allow quantification of expression.

In a preferred embodiment, gene expression or protein monitoring of a number of entities, i.e., an expression profile, is monitored simultaneously. Such profiles will typically involve a plurality of those entities described herein..

In this embodiment, the breast cancer nucleic acid probes are attached to biochips as outlined herein for the detection and quantification of breast cancer sequences in a particular cell. Alternatively, PCR may be used. Thus, a series, e.g., of microtiter plate,

may be used with dispensed primers in desired wells. A PCR reaction can then be performed and analyzed for each well.

Expression monitoring can be performed to identify compounds that modify the expression of one or more breast cancer-associated sequences, e.g., a polynucleotide sequence set out inTable 1. Generally, in a preferred embodiment, a test modulator is added to the cells prior to analysis. Moreover, screens are also provided to identify agents that modulate breast cancer, modulate breast cancer proteins, bind to a breast cancer protein, or interfere with the binding of a breast cancer protein and an antibody or other binding partner.

The term "test compound" or "drug candidate" or "modulator" or grammatical equivalents as used herein describes any molecule, e.g., protein, oligopeptide, small organic molecule, polysaccharide, polynucleotide, etc., to be tested for the capacity to directly or indirectly alter the breast cancer phenotype or the expression of a breast cancer sequence, e.g., a nucleic acid or protein sequence. In preferred embodiments, modulators alter expression profiles, or expression profile nucleic acids or proteins provided herein. In one embodiment, the modulator suppresses a breast cancer phenotype, e.g. to a normal tissue fingerprint. In another embodiment, a modulator induced a breast cancer phenotype. Generally, a plurality of assay mixtures are run in parallel with different agent concentrations to obtain a differential response to the various concentrations. Typically, one of these concentrations serves as a negative control, i.e., at zero concentration or below the level of detection.

Drug candidates encompass numerous chemical classes, though typically they are organic molecules, preferably small organic compounds having a molecular weight of more than 100 and less than about 2,500 daltons. Preferred small molecules are less than 2000, or less than 1500 or less than 1000 or less than 500 D. Candidate agents comprise functional groups necessary for structural interaction with proteins, particularly hydrogen bonding, and typically include at least an amine, carbonyl, hydroxyl or carboxyl group, preferably at least two of the functional chemical groups. The candidate agents often comprise cyclical carbon or heterocyclic structures and/or aromatic or polyaromatic structures substituted with one or more of the above functional groups. Candidate agents are also found among biomolecules including peptides, saccharides, fatty acids, steroids, purines, pyrimidines, derivatives, structural analogs or combinations thereof. Particularly preferred are peptides.

In one aspect, a modulator will neutralize the effect of a breast cancer protein. By "neutralize" is meant that activity of a protein is inhibited or blocked and the consequent effect on the cell.

In certain embodiments, combinatorial libraries of potential modulators will be screened for an ability to bind to a breast cancer polypeptide or to modulate activity. Conventionally, new chemical entities with useful properties are generated by identifying a chemical compound (called a "lead compound") with some desirable property or activity, e.g., inhibiting activity, creating variants of the lead compound, and evaluating the property and activity of those variant compounds. Often, high throughput screening (HTS) methods are employed for such an analysis.

In one preferred embodiment, high throughput screening methods involve providing a library containing a large number of potential therapeutic compounds (candidate compounds). Such "combinatorial chemical libraries" are then screened in one or more assays to identify those library members (particular chemical species or subclasses) that display a desired characteristic activity. The compounds thus identified can serve as conventional "lead compounds" or can themselves be used as potential or actual therapeutics.

A combinatorial chemical library is a collection of diverse chemical compounds generated by either chemical synthesis or biological synthesis by combining a number of chemical "building blocks" such as reagents. For example, a linear combinatorial chemical library, such as a polypeptide (e.g., mutein) library, is formed by combining a set of chemical building blocks called amino acids in every possible way for a given compound length (i.e., the number of amino acids in a polypeptide compound). Millions of chemical compounds can be synthesized through such combinatorial mixing of chemical building blocks (Gallop *et al., J. Med. Chem.* 37(9):1233-1251 (1994)).

Preparation and screening of combinatorial chemical libraries is well known to those of skill in the art. Such combinatorial chemical libraries include, but are not limited to, peptide libraries (*see, e.g.,* U.S. Patent No. 5,010,175, Furka, *Pept. Prot. Res.* 37:487-493 (1991), Houghton *et al.*, *Nature*, 354:84-88 (1991)), peptoids (PCT Publication No WO 91/19735), encoded peptides (PCT Publication WO 93/20242), random bio-oligomers (PCT Publication WO 92/00091), benzodiazepines (U.S. Pat. No. 5,288,514), diversomers such as hydantoins, benzodiazepines and dipeptides (Hobbs *et al., Proc. Nat. Acad. Sci. USA* 90:6909-6913 (1993)), vinylogous polypeptides (Hagihara *et al., J. Amer. Chem. Soc.*

114:6568 (1992)), nonpeptidal peptidomimetics with a Beta-D-Glucose scaffolding (Hirschmann *et al., J. Amer. Chem. Soc.* 114:9217-9218 (1992)), analogous organic syntheses of small compound libraries (Chen *et al., J. Amer. Chem.* Soc. 116:2661 (1994)), oligocarbamates (Cho, *et al., Science* 261:1303 (1993)), and/or peptidyl phosphonates (Campbell *et al., J. Org. Chem.* 59:658 (1994)). *See, generally,* Gordon *et al., J. Med. Chem.* 37:1385 (1994), nucleic acid libraries (*see, e.g.,* Strategene, Corp.), peptide nucleic acid libraries (*see, e.g.,* U.S. Patent 5,539,083), antibody libraries (*see, e.g.,* Vaughn *et al., Nature Biotechnology* 14(3):309-314 (1996), and PCT/US96/10287), carbohydrate libraries (*see, e.g.,* Liang *et al., Science* 274:1520-1522 (1996), and U.S. Patent No. 5,593,853), and small organic molecule libraries (*see, e.g.,* benzodiazepines, Baum, C&EN, Jan 18, page 33 (1993); isoprenoids, U.S. Patent No. 5,569,588; thiazolidinones and metathiazanones, U.S. Patent No. 5,549,974; pyrrolidines, U.S. Patent Nos. 5,525,735 and 5,519,134; morpholino compounds, U.S. Patent No. 5,506,337; benzodiazepines, U.S. Patent No. 5,288,514; and the like).

Devices for the preparation of combinatorial libraries are commercially available (*see, e.g.,* 357 MPS, 390 MPS, Advanced Chem Tech, Louisville KY, Symphony, Rainin, Woburn, MA, 433A Applied Biosystems, Foster City, CA, 9050 Plus, Millipore, Bedford, MA).

A number of well known robotic systems have also been developed for solution phase chemistries. These systems include automated workstations like the automated synthesis apparatus developed by Takeda Chemical Industries, LTD. (Osaka, Japan) and many robotic systems utilizing robotic arms (Zymate II, Zymark Corporation, Hopkinton, Mass.; Orca, Hewlett-Packard, Palo Alto, Calif.), which mimic the manual synthetic operations performed by a chemist. Any of the above devices are suitable for use with the present invention. The nature and implementation of modifications to these devices (if any) so that they can operate as discussed herein will be apparent to persons skilled in the relevant art. In addition, numerous combinatorial libraries are themselves commercially available (*see, e.g.,* ComGenex, Princeton, N.J., Asinex, Moscow, Ru, Tripos, Inc., St. Louis, MO, ChemStar, Ltd, Moscow, RU, 3D Pharmaceuticals, Exton, PA, Martek Biosciences, Columbia, MD, *etc.*).

The assays to identify modulators are amenable to high throughput screening. Preferred assays thus detect enhancement or inhibition of breast cancer gene transcription,

inhibition or enhancement of polypeptide expression, and inhibition or enhancement of polypeptide activity.

High throughput assays for the presence, absence, quantification, or other properties of particular nucleic acids or protein products are well known to those of skill in the art. Similarly, binding assays and reporter gene assays are similarly well known. Thus, e.g., U.S. Patent No. 5,559,410 discloses high throughput screening methods for proteins, U.S. Patent No. 5,585,639 discloses high throughput screening methods for nucleic acid binding (i.e., in arrays), while U.S. Patent Nos. 5,576,220 and 5,541,061 disclose high throughput methods of screening for ligand/antibody binding.

In addition, high throughput screening systems are commercially available (*see, e.g.,* Zymark Corp., Hopkinton, MA; Air Technical Industries, Mentor, OH; Beckman Instruments, Inc. Fullerton, CA; Precision Systems, Inc., Natick, MA, *etc.*). These systems typically automate entire procedures, including all sample and reagent pipetting, liquid dispensing, timed incubations, and final readings of the microplate in detector(s) appropriate for the assay. These configurable systems provide high throughput and rapid start up as well as a high degree of flexibility and customization. The manufacturers of such systems provide detailed protocols for various high throughput systems. Thus, e.g., Zymark Corp. provides technical bulletins describing screening systems for detecting the modulation of gene transcription, ligand binding, and the like.

In one embodiment, modulators are proteins, often naturally occurring proteins or fragments of naturally occurring proteins. Thus, *e.g.*, cellular extracts containing proteins, or random or directed digests of proteinaceous cellular extracts, may be used. In this way libraries of proteins may be made for screening in the methods of the invention. Particularly preferred in this embodiment are libraries of bacterial, fungal, viral, and mammalian proteins, with the latter being preferred, and human proteins being especially preferred. Particularly useful test compound will be directed to the class of proteins to which the target belongs, *e.g.*, substrates for enzymes or ligands and receptors.

In a preferred embodiment, modulators are peptides of from about 5 to about 30 amino acids, with from about 5 to about 20 amino acids being preferred, and from about 7 to about 15 being particularly preferred. The peptides may be digests of naturally occurring proteins as is outlined above, random peptides, or "biased" random peptides. By "randomized" or grammatical equivalents herein is meant that each nucleic acid and peptide

consists of essentially random nucleotides and amino acids, respectively. Since generally these random peptides (or nucleic acids, discussed below) are chemically synthesized, they may incorporate any nucleotide or amino acid at any position. The synthetic process can be designed to generate randomized proteins or nucleic acids, to allow the formation of all or most of the possible combinations over the length of the sequence, thus forming a library of randomized candidate bioactive proteinaceous agents.

In one embodiment, the library is fully randomized, with no sequence preferences or constants at any position. In a preferred embodiment, the library is biased. That is, some positions within the sequence are either held constant, or are selected from a limited number of possibilities. For example, in a preferred embodiment, the nucleotides or amino acid residues are randomized within a defined class, e.g., of hydrophobic amino acids, hydrophilic residues, sterically biased (either small or large) residues, towards the creation of nucleic acid binding domains, the creation of cysteines, for cross-linking, prolines for SH-3 domains, serines, threonines, tyrosines or histidines for phosphorylation sites, etc., or to purines, etc.

Modulators of breast cancer can also be nucleic acids, as defined above.

As described above generally for proteins, nucleic acid modulating agents may be naturally occurring nucleic acids, random nucleic acids, or "biased" random nucleic acids. For example, digests of procaryotic or eucaryotic genomes may be used as is outlined above for proteins.

In a preferred embodiment, the candidate compounds are organic chemical moieties, a wide variety of which are available in the literature.

After the candidate agent has been added and the cells allowed to incubate for some period of time, the sample containing a target sequence to be analyzed is added to the biochip. If required, the target sequence is prepared using known techniques. For example, the sample may be treated to lyse the cells, using known lysis buffers, electroporation, etc., with purification and/or amplification such as PCR performed as appropriate. For example, an *in vitro* transcription with labels covalently attached to the nucleotides is performed. Generally, the nucleic acids are labeled with biotin-FITC or PE, or with cy3 or cy5.

In a preferred embodiment, the target sequence is labeled with, e.g., a fluorescent, a chemiluminescent, a chemical, or a radioactive signal, to provide a means of detecting the target sequence's specific binding to a probe. The label also can be an enzyme,

such as, alkaline phosphatase or horseradish peroxidase, which when provided with an appropriate substrate produces a product that can be detected. Alternatively, the label can be a labeled compound or small molecule, such as an enzyme inhibitor, that binds but is not catalyzed or altered by the enzyme. The label also can be a moiety or compound, such as, an epitope tag or biotin which specifically binds to streptavidin. For the example of biotin, the streptavidin is labeled as described above, thereby, providing a detectable signal for the bound target sequence. Unbound labeled streptavidin is typically removed prior to analysis.

As will be appreciated by those in the art, these assays can be direct hybridization assays or can comprise "sandwich assays", which include the use of multiple probes, as is generally outlined in U.S. Patent Nos. 5,681,702, 5,597,909, 5,545,730, 5,594,117, 5,591,584, 5,571,670, 5,580,731, 5,571,670, 5,591,584, 5,624,802, 5,635,352, 5,594,118, 5,359,100, 5,124,246 and 5,681,697, all of which are hereby incorporated by reference. In this embodiment, in general, the target nucleic acid is prepared as outlined above, and then added to the biochip comprising a plurality of nucleic acid probes, under conditions that allow the formation of a hybridization complex.

A variety of hybridization conditions may be used in the present invention, including high, moderate and low stringency conditions as outlined above. The assays are generally run under stringency conditions which allows formation of the label probe hybridization complex only in the presence of target. Stringency can be controlled by altering a step parameter that is a thermodynamic variable, including, but not limited to, temperature, formamide concentration, salt concentration, chaotropic salt concentration pH, organic solvent concentration, etc.

These parameters may also be used to control non-specific binding, as is generally outlined in U.S. Patent No. 5,681,697. Thus it may be desirable to perform certain steps at higher stringency conditions to reduce non-specific binding.

The reactions outlined herein may be accomplished in a variety of ways. Components of the reaction may be added simultaneously, or sequentially, in different orders, with preferred embodiments outlined below. In addition, the reaction may include a variety of other reagents. These include salts, buffers, neutral proteins, e.g. albumin, detergents, *etc.* which may be used to facilitate optimal hybridization and detection, and/or reduce non-specific or background interactions. Reagents that otherwise improve the efficiency of the

assay, such as protease inhibitors, nuclease inhibitors, anti-microbial agents, *etc.*, may also be used as appropriate, depending on the sample preparation methods and purity of the target.

The assay data are analyzed to determine the expression levels, and changes in expression levels as between states, of individual genes, forming a gene expression profile.

Screens are performed to identify modulators of the breast cancer phenotype. In one embodiment, screening is performed to identify modulators that can induce or suppress a particular expression profile, thus preferably generating the associated phenotype. In another embodiment, *e.g.*, for diagnostic applications, having identified differentially expressed genes important in a particular state, screens can be performed to identify modulators that alter expression of individual genes. In an another embodiment, screening is performed to identify modulators that alter a biological function of the expression product of a differentially expressed gene. Again, having identified the importance of a gene in a particular state, screens are performed to identify agents that bind and/or modulate the biological activity of the gene product.

In addition screens can be done for genes that are induced in response to a candidate agent. After identifying a modulator based upon its ability to suppress a breast cancer expression pattern leading to a normal expression pattern, or to modulate a single breast cancer gene expression profile so as to mimic the expression of the gene from normal tissue, a screen as described above can be performed to identify genes that are specifically modulated in response to the agent. Comparing expression profiles between normal tissue and agent treated breast cancer tissue reveals genes that are not expressed in normal tissue or breast cancer tissue, but are expressed in agent treated tissue. These agent-specific sequences can be identified and used by methods described herein for breast cancer genes or proteins. In particular these sequences and the proteins they encode find use in marking or identifying agent treated cells. In addition, antibodies can be raised against the agent induced proteins and used to target novel therapeutics to the treated breast cancer tissue sample.

Thus, in one embodiment, a test compound is administered to a population of breast cancer cells, that have an associated breast cancer expression profile. By "administration" or "contacting" herein is meant that the candidate agent is added to the cells in such a manner as to allow the agent to act upon the cell, whether by uptake and intracellular action, or by action at the cell surface. In some embodiments, nucleic acid encoding a proteinaceous candidate agent (i.e., a peptide) may be put into a viral construct

such as an adenoviral or retroviral construct, and added to the cell, such that expression of the peptide agent is accomplished, e.g., PCT US97/01019. Regulatable gene therapy systems can also be used.

Once the test compound has been administered to the cells, the cells can be washed if desired and are allowed to incubate under preferably physiological conditions for some period of time. The cells are then harvested and a new gene expression profile is generated, as outlined herein.

Thus, e.g., breast cancer tissue may be screened for agents that modulate, e.g., induce or suppress the breast cancer phenotype. A change in at least one gene, preferably many, of the expression profile indicates that the agent has an effect on breast cancer activity. By defining such a signature for the breast cancer phenotype, screens for new drugs that alter the phenotype can be devised. With this approach, the drug target need not be known and need not be represented in the original expression screening platform, nor does the level of transcript for the target protein need to change.

In a preferred embodiment, as outlined above, screens may be done on individual genes and gene products (proteins). That is, having identified a particular differentially expressed gene as important in a particular state, screening of modulators of either the expression of the gene or the gene product itself can be done. The gene products of differentially expressed genes are sometimes referred to herein as "breast cancer proteins" or a "breast cancer modulatory protein". The breast cancer modulatory protein may be a fragment, or alternatively, be the full length protein to the fragment encoded by the nucleic acids of the Tables. Preferably, the breast cancer modulatory protein is a fragment. In a preferred embodiment, the breast cancer amino acid sequence which is used to determine sequence identity or similarity is encoded by a nucleic acid of Table 2. In another embodiment, the sequences are naturally occurring allelic variants of a protein encoded by a nucleic acid of Table 2. In another embodiment, the sequences are sequence variants as further described herein.

Preferably, the breast cancer modulatory protein is a fragment of approximately 14 to 24 amino acids long. More preferably the fragment is a soluble fragment. Preferably, the fragment includes a non-transmembrane region. In a preferred embodiment, the fragment has an N-terminal Cys to aid in solubility. In one embodiment, the

C-terminus of the fragment is kept as a free acid and the N-terminus is a free amine to aid in coupling, i.e., to cysteine.

In one embodiment the breast cancer proteins are conjugated to an immunogenic agent as discussed herein. In one embodiment the breast cancer protein is conjugated to BSA.

Measurements of breast cancer polypeptide activity, or of breast cancer or the breast cancer phenotype can be performed using a variety of assays. For example, the effects of the test compounds upon the function of the breast cancer polypeptides can be measured by examining parameters described above. A suitable physiological change that affects activity can be used to assess the influence of a test compound on the polypeptides of this invention. When the functional consequences are determined using intact cells or animals, one can also measure a variety of effects such as, in the case of breast cancer associated with tumors, tumor growth, tumor metastasis, neovascularization, hormone release, transcriptional changes to both known and uncharacterized genetic markers (e.g., northern blots), changes in cell metabolism such as cell growth or pH changes, and changes in intracellular second messengers such as cGMP. In the assays of the invention, mammalian breast cancer polypeptide is typically used, e.g., mouse, preferably human.

Assays to identify compounds with modulating activity can be performed *in vitro*. For example, a breast cancer polypeptide is first contacted with a potential modulator and incubated for a suitable amount of time, e.g., from 0.5 to 48 hours. In one embodiment, the breast cancer polypeptide levels are determined *in vitro* by measuring the level of protein or mRNA. The level of protein is measured using immunoassays such as western blotting, ELISA and the like with an antibody that selectively binds to the breast cancer polypeptide or a fragment thereof. For measurement of mRNA, amplification, e.g., using PCR, LCR, or hybridization assays, e.g., northern hybridization, RNAse protection, dot blotting, are preferred. The level of protein or mRNA is detected using directly or indirectly labeled detection agents, e.g., fluorescently or radioactively labeled nucleic acids, radioactively or enzymatically labeled antibodies, and the like, as described herein.

Alternatively, a reporter gene system can be devised using the breast cancer protein promoter operably linked to a reporter gene such as luciferase, green fluorescent protein, CAT, or β-gal. The reporter construct is typically transfected into a cell. After

treatment with a potential modulator, the amount of reporter gene transcription, translation, or activity is measured according to standard techniques known to those of skill in the art.

In a preferred embodiment, as outlined above, screens may be done on individual genes and gene products (proteins). That is, having identified a particular differentially expressed gene as important in a particular state, screening of modulators of the expression of the gene or the gene product itself can be done. The gene products of differentially expressed genes are sometimes referred to herein as "breast cancer proteins." The breast cancer protein may be a fragment, or alternatively, be the full length protein to a fragment shown herein.

In one embodiment, screening for modulators of expression of specific genes is performed. Typically, the expression of only one or a few genes are evaluated. In another embodiment, screens are designed to first find compounds that bind to differentially expressed proteins. These compounds are then evaluated for the ability to modulate differentially expressed activity. Moreover, once initial candidate compounds are identified, variants can be further screened to better evaluate structure activity relationships.

In a preferred embodiment, binding assays are done. In general, purified or isolated gene product is used; that is, the gene products of one or more differentially expressed nucleic acids are made. For example, antibodies are generated to the protein gene products, and standard immunoassays are run to determine the amount of protein present. Alternatively, cells comprising the breast cancer proteins can be used in the assays.

Thus, in a preferred embodiment, the methods comprise combining a breast cancer protein and a candidate compound, and determining the binding of the compound to the breast cancer protein. Preferred embodiments utilize the human breast cancer protein, although other mammalian proteins may also be used, e.g. for the development of animal models of human disease. In some embodiments, as outlined herein, variant or derivative breast cancer proteins may be used.

Generally, in a preferred embodiment of the methods herein, the breast cancer protein or the candidate agent is non-diffusably bound to an insoluble support having isolated sample receiving areas (e.g. a microtiter plate, an array, etc.). The insoluble supports may be made of any composition to which the compositions can be bound, is readily separated from soluble material, and is otherwise compatible with the overall method of screening. The surface of such supports may be solid or porous and of any convenient shape. Examples of

suitable insoluble supports include microtiter plates, arrays, membranes and beads. These are typically made of glass, plastic (e.g., polystyrene), polysaccharides, nylon or nitrocellulose, teflon™, etc. Microtiter plates and arrays are especially convenient because a large number of assays can be carried out simultaneously, using small amounts of reagents and samples. The particular manner of binding of the composition is not crucial so long as it is compatible with the reagents and overall methods of the invention, maintains the activity of the composition and is nondiffusable. Preferred methods of binding include the use of antibodies (which do not sterically block either the ligand binding site or activation sequence when the protein is bound to the support), direct binding to "sticky" or ionic supports, chemical crosslinking, the synthesis of the protein or agent on the surface, etc. Following binding of the protein or agent, excess unbound material is removed by washing. The sample receiving areas may then be blocked through incubation with bovine serum albumin (BSA), casein or other innocuous protein or other moiety.

In a preferred embodiment, the breast cancer protein is bound to the support, and a test compound is added to the assay. Alternatively, the candidate agent is bound to the support and the breast cancer protein is added. Novel binding agents include specific antibodies, non-natural binding agents identified in screens of chemical libraries, peptide analogs, etc. Of particular interest are screening assays for agents that have a low toxicity for human cells. A wide variety of assays may be used for this purpose, including labeled in vitro protein-protein binding assays, electrophoretic mobility shift assays, immunoassays for protein binding, functional assays (phosphorylation assays, etc.) and the like.

The determination of the binding of the test modulating compound to the breast cancer protein may be done in a number of ways. In a preferred embodiment, the compound is labeled, and binding determined directly, e.g., by attaching all or a portion of the breast cancer protein to a solid support, adding a labeled candidate agent (e.g., a fluorescent label), washing off excess reagent, and determining whether the label is present on the solid support. Various blocking and washing steps may be utilized as appropriate.

In some embodiments, only one of the components is labeled, e.g., the proteins (or proteinaceous candidate compounds) can be labeled. Alternatively, more than one component can be labeled with different labels, e.g., ^{125}I for the proteins and a fluorophor for the compound. Proximity reagents, e.g., quenching or energy transfer reagents are also useful.

In one embodiment, the binding of the test compound is determined by competitive binding assay. The competitor is a binding moiety known to bind to the target molecule (i.e., a breast cancer protein), such as an antibody, peptide, binding partner, ligand, etc. Under certain circumstances, there may be competitive binding between the compound and the binding moiety, with the binding moiety displacing the compound. In one embodiment, the test compound is labeled. Either the compound, or the competitor, or both, is added first to the protein for a time sufficient to allow binding, if present. Incubations may be performed at a temperature which facilitates optimal activity, typically between 4 and 40°C. Incubation periods are typically optimized, e.g., to facilitate rapid high throughput screening. Typically between 0.1 and 1 hour will be sufficient. Excess reagent is generally removed or washed away. The second component is then added, and the presence or absence of the labeled component is followed, to indicate binding.

In a preferred embodiment, the competitor is added first, followed by the test compound. Displacement of the competitor is an indication that the test compound is binding to the breast cancer protein and thus is capable of binding to, and potentially modulating, the activity of the breast cancer protein. In this embodiment, either component can be labeled. Thus, e.g., if the competitor is labeled, the presence of label in the wash solution indicates displacement by the agent. Alternatively, if the test compound is labeled, the presence of the label on the support indicates displacement.

In an alternative embodiment, the test compound is added first, with incubation and washing, followed by the competitor. The absence of binding by the competitor may indicate that the test compound is bound to the breast cancer protein with a higher affinity. Thus, if the test compound is labeled, the presence of the label on the support, coupled with a lack of competitor binding, may indicate that the test compound is capable of binding to the breast cancer protein.

In a preferred embodiment, the methods comprise differential screening to identity agents that are capable of modulating the activity of the breast cancer proteins. In this embodiment, the methods comprise combining a breast cancer protein and a competitor in a first sample. A second sample comprises a test compound, a breast cancer protein, and a competitor. The binding of the competitor is determined for both samples, and a change, or difference in binding between the two samples indicates the presence of an agent capable of binding to the breast cancer protein and potentially modulating its activity. That is, if the

binding of the competitor is different in the second sample relative to the first sample, the agent is capable of binding to the breast cancer protein.

Alternatively, differential screening is used to identify drug candidates that bind to the native breast cancer protein, but cannot bind to modified breast cancer proteins. The structure of the breast cancer protein may be modeled, and used in rational drug design to synthesize agents that interact with that site. Drug candidates that affect the activity of a breast cancer protein are also identified by screening drugs for the ability to either enhance or reduce the activity of the protein.

Positive controls and negative controls may be used in the assays. Preferably control and test samples are performed in at least triplicate to obtain statistically significant results. Incubation of all samples is for a time sufficient for the binding of the agent to the protein. Following incubation, samples are washed free of non-specifically bound material and the amount of bound, generally labeled agent determined. For example, where a radiolabel is employed, the samples may be counted in a scintillation counter to determine the amount of bound compound.

A variety of other reagents may be included in the screening assays. These include reagents like salts, neutral proteins, e.g. albumin, detergents, etc. which may be used to facilitate optimal protein-protein binding and/or reduce non-specific or background interactions. Also reagents that otherwise improve the efficiency of the assay, such as protease inhibitors, nuclease inhibitors, anti-microbial agents, etc., may be used. The mixture of components may be added in an order that provides for the requisite binding.

In a preferred embodiment, the invention provides methods for screening for a compound capable of modulating the activity of a breast cancer protein. The methods comprise adding a test compound, as defined above, to a cell comprising breast cancer proteins. Preferred cell types include almost any cell. The cells contain a recombinant nucleic acid that encodes a breast cancer protein. In a preferred embodiment, a library of candidate agents are tested on a plurality of cells.

In one aspect, the assays are evaluated in the presence or absence or previous or subsequent exposure of physiological signals, e.g. hormones, antibodies, peptides, antigens, cytokines, growth factors, action potentials, pharmacological agents including chemotherapeutics, radiation, carcinogenics, or other cells (i.e. cell-cell contacts). In another example, the determinations are determined at different stages of the cell cycle process.

In this way, compounds that modulate breast cancer agents are identified. Compounds with pharmacological activity are able to enhance or interfere with the activity of the breast cancer protein. Once identified, similar structures are evaluated to identify critical structural feature of the compound.

In one embodiment, a method of inhibiting breast cancer cell division is provided. The method comprises administration of a breast cancer inhibitor. In another embodiment, a method of inhibiting breast cancer is provided. The method comprises administration of a breast cancer inhibitor. In a further embodiment, methods of treating cells or individuals with breast cancer are provided. The method comprises administration of a breast cancer inhibitor.

In one embodiment, a breast cancer inhibitor is an antibody as discussed above. In another embodiment, the breast cancer inhibitor is an antisense molecule.

A variety of cell growth, proliferation, and metastasis assays are known to those of skill in the art, as described below.

Soft agar growth or colony formation in suspension

Normal cells require a solid substrate to attach and grow. When the cells are transformed, they lose this phenotype and grow detached from the substrate. For example, transformed cells can grow in stirred suspension culture or suspended in semi-solid media, such as semi-solid or soft agar. The transformed cells, when transfected with tumor suppressor genes, regenerate normal phenotype and require a solid substrate to attach and grow. Soft agar growth or colony formation in suspension assays can be used to identify modulators of breast cancer sequences, which when expressed in host cells, inhibit abnormal cellular proliferation and transformation. A therapeutic compound would reduce or eliminate the host cells' ability to grow in stirred suspension culture or suspended in semi-solid media, such as semi-solid or soft.

Techniques for soft agar growth or colony formation in suspension assays are described in Freshney, *Culture of Animal Cells a Manual of Basic Technique* (3rd ed., 1994), herein incorporated by reference. *See also*, the methods section of Garkavtsev *et al.* (1996), *supra*, herein incorporated by reference.

Contact inhibition and density limitation of growth

Normal cells typically grow in a flat and organized pattern in a petri dish until they touch other cells. When the cells touch one another, they are contact inhibited and stop

growing. When cells are transformed, however, the cells are not contact inhibited and continue to grow to high densities in disorganized foci. Thus, the transformed cells grow to a higher saturation density than normal cells. This can be detected morphologically by the formation of a disoriented monolayer of cells or rounded cells in foci within the regular pattern of normal surrounding cells. Alternatively, labeling index with (^{3}H)-thymidine at saturation density can be used to measure density limitation of growth. *See* Freshney (1994), *supra*. The transformed cells, when transfected with tumor suppressor genes, regenerate a normal phenotype and become contact inhibited and would grow to a lower density.

In this assay, labeling index with (^{3}H)-thymidine at saturation density is a preferred method of measuring density limitation of growth. Transformed host cells are transfected with a breast cancer-associated sequence and are grown for 24 hours at saturation density in non-limiting medium conditions. The percentage of cells labeling with (^{3}H)-thymidine is determined autoradiographically. *See*, Freshney (1994), *supra*.

Growth factor or serum dependence

Transformed cells have a lower serum dependence than their normal counterparts (*see, e.g.*, Temin, *J. Natl. Cancer Insti.* 37:167-175 (1966); Eagle *et al.*, *J. Exp. Med.* 131:836-879 (1970)); Freshney, *supra*. This is in part due to release of various growth factors by the transformed cells. Growth factor or serum dependence of transformed host cells can be compared with that of control.

Tumor specific markers levels

Tumor cells release an increased amount of certain factors (hereinafter "tumor specific markers") than their normal counterparts. For example, plasminogen activator (PA) is released from human glioma at a higher level than from normal brain cells (*see, e.g.*, Gullino, *Angiogenesis, tumor vascularization, and potential interference with tumor growth.* in *Biological Responses in Cancer*, pp. 178-184 (Mihich (ed.) 1985)). Similarly, Tumor angiogenesis factor (TAF) is released at a higher level in tumor cells than their normal counterparts. *See, e.g.*, Folkman, *Angiogenesis and Cancer, Sem Cancer Biol.* (1992)).

Various techniques which measure the release of these factors are described in Freshney (1994), *supra*. Also, *see*, Unkless *et al.* , *J. Biol. Chem.* 249:4295-4305 (1974); Strickland & Beers, *J. Biol. Chem.* 251:5694-5702 (1976); Whur *et al.*, *Br. J. Cancer* 42:305-

312 (1980); Gullino, *Angiogenesis, tumor vascularization, and potential interference with tumor growth.* in *Biological Responses in Cancer*, pp. 178-184 (Mihich (ed.) 1985); Freshney *Anticancer Res.* 5:111-130 (1985).

Invasiveness into Matrigel

The degree of invasiveness into Matrigel or some other extracellular matrix constituent can be used as an assay to identify compounds that modulate breast cancer-associated sequences. Tumor cells exhibit a good correlation between malignancy and invasiveness of cells into Matrigel or some other extracellular matrix constituent. In this assay, tumorigenic cells are typically used as host cells. Expression of a tumor suppressor gene in these host cells would decrease invasiveness of the host cells.

Techniques described in Freshney (1994), *supra*, can be used. Briefly, the level of invasion of host cells can be measured by using filters coated with Matrigel or some other extracellular matrix constituent. Penetration into the gel, or through to the distal side of the filter, is rated as invasiveness, and rated histologically by number of cells and distance moved, or by prelabeling the cells with ^{125}I and counting the radioactivity on the distal side of the filter or bottom of the dish. *See, e.g.,* Freshney (1984), *supra*.

Tumor growth in vivo

Effects of breast cancer-associated sequences on cell growth can be tested in transgenic or immune-suppressed mice. Knock-out transgenic mice can be made, in which the breast cancer gene is disrupted or in which a breast cancer gene is inserted. Knock-out transgenic mice can be made by insertion of a marker gene or other heterologous gene into the endogenous breast cancer gene site in the mouse genome via homologous recombination. Such mice can also be made by substituting the endogenous breast cancer gene with a mutated version of the breast cancer gene, or by mutating the endogenous breast cancer gene, e.g., by exposure to carcinogens.

A DNA construct is introduced into the nuclei of embryonic stem cells. Cells containing the newly engineered genetic lesion are injected into a host mouse embryo, which is re-implanted into a recipient female. Some of these embryos develop into chimeric mice that possess germ cells partially derived from the mutant cell line. Therefore, by breeding the chimeric mice it is possible to obtain a new line of mice containing the introduced genetic

lesion (*see, e.g.*, Capecchi *et al., Science* 244:1288 (1989)). Chimeric targeted mice can be derived according to Hogan *et al.*, *Manipulating the Mouse Embryo*: *A Laboratory Manual*, Cold Spring Harbor Laboratory (1988) and *Teratocarcinomas and Embryonic Stem Cells: A Practical Approach*, Robertson, ed., IRL Press, Washington, D.C., (1987).

Alternatively, various immune-suppressed or immune-deficient host animals can be used. For example, genetically athymic "nude" mouse (*see, e.g.*, Giovanella *et al.*, *J. Natl. Cancer Inst.* 52:921 (1974)), a SCID mouse, a thymectomized mouse, or an irradiated mouse (*see, e.g.*, Bradley *et al.*, *Br. J. Cancer* 38:263 (1978); Selby *et al.*, *Br. J. Cancer* 41:52 (1980)) can be used as a host. Transplantable tumor cells (typically about 10^6 cells) injected into isogenic hosts will produce invasive tumors in a high proportions of cases, while normal cells of similar origin will not. In hosts which developed invasive tumors, cells expressing a breast cancer-associated sequences are injected subcutaneously. After a suitable length of time, preferably 4-8 weeks, tumor growth is measured (e.g., by volume or by its two largest dimensions) and compared to the control. Tumors that have statistically significant reduction (using, e.g., Student's T test) are said to have inhibited growth.

Polynucleotide modulators of breast cancer

Antisense Polynucleotides

In certain embodiments, the activity of a breast cancer-associated protein is down-regulated, or entirely inhibited, by the use of antisense polynucleotide, *i.e.*, a nucleic acid complementary to, and which can preferably hybridize specifically to, a coding mRNA nucleic acid sequence, e.g., a breast cancer protein mRNA, or a subsequence thereof. Binding of the antisense polynucleotide to the mRNA reduces the translation and/or stability of the mRNA.

In the context of this invention, antisense polynucleotides can comprise naturally-occurring nucleotides, or synthetic species formed from naturally-occurring subunits or their close homologs. Antisense polynucleotides may also have altered sugar moieties or inter-sugar linkages. Exemplary among these are the phosphorothioate and other sulfur containing species which are known for use in the art. Analogs are comprehended by this invention so long as they function effectively to hybridize with the breast cancer protein mRNA. *See, e.g.*, Isis Pharmaceuticals, Carlsbad, CA; Sequitor, Inc., Natick, MA.

Such antisense polynucleotides can readily be synthesized using recombinant means, or can be synthesized *in vitro*. Equipment for such synthesis is sold by several vendors, including Applied Biosystems. The preparation of other oligonucleotides such as phosphorothioates and alkylated derivatives is also well known to those of skill in the art.

Antisense molecules as used herein include antisense or sense oligonucleotides. Sense oligonucleotides can, e.g., be employed to block transcription by binding to the anti-sense strand. The antisense and sense oligonucleotide comprise a single-stranded nucleic acid sequence (either RNA or DNA) capable of binding to target mRNA (sense) or DNA (antisense) sequences for breast cancer molecules. A preferred antisense molecule is for a breast cancer sequences in Table 1, or for a ligand or activator thereof. Antisense or sense oligonucleotides, according to the present invention, comprise a fragment generally at least about 14 nucleotides, preferably from about 14 to 30 nucleotides. The ability to derive an antisense or a sense oligonucleotide, based upon a cDNA sequence encoding a given protein is described in, e.g., Stein & Cohen (*Cancer Res.* 48:2659 (1988 and van der Krol *et al.* (*BioTechniques* 6:958 (1988)).

Ribozymes

In addition to antisense polynucleotides, ribozymes can be used to target and inhibit transcription of breast cancer-associated nucleotide sequences. A ribozyme is an RNA molecule that catalytically cleaves other RNA molecules. Different kinds of ribozymes have been described, including group I ribozymes, hammerhead ribozymes, hairpin ribozymes, RNase P, and axhead ribozymes (*see, e.g.,* Castanotto *et al.*, *Adv. in Pharmacology* 25: 289-317 (1994) for a general review of the properties of different ribozymes).

The general features of hairpin ribozymes are described, e.g., in Hampel *et al.*, *Nucl. Acids Res.* 18:299-304 (1990); European Patent Publication No. 0 360 257; U.S. Patent No. 5,254,678. Methods of preparing are well known to those of skill in the art (*see, e.g.,* WO 94/26877; Ojwang *et al.*, *Proc. Natl. Acad. Sci.* USA 90:6340-6344 (1993); Yamada *et al.*, *Human Gene Therapy* 1:39-45 (1994); Leavitt *et al.*, *Proc. Natl. Acad. Sci. USA* 92:699-703 (1995); Leavitt *et al.*, *Human Gene Therapy* 5:1151-120 (1994); and Yamada *et al.*, *Virology* 205: 121-126 (1994)).

Polynucleotide modulators of breast cancer may be introduced into a cell containing the target nucleotide sequence by formation of a conjugate with a ligand binding molecule, as described in WO 91/04753. Suitable ligand binding molecules include, but are not limited to, cell surface receptors, growth factors, other cytokines, or other ligands that bind to cell surface receptors. Preferably, conjugation of the ligand binding molecule does not substantially interfere with the ability of the ligand binding molecule to bind to its corresponding molecule or receptor, or block entry of the sense or antisense oligonucleotide or its conjugated version into the cell. Alternatively, a polynucleotide modulator of breast cancer may be introduced into a cell containing the target nucleic acid sequence, e.g., by formation of an polynucleotide-lipid complex, as described in WO 90/10448. It is understood that the use of antisense molecules or knock out and knock in models may also be used in screening assays as discussed above, in addition to methods of treatment.

Thus, in one embodiment, methods of modulating breast cancer in cells or organisms are provided. In one embodiment, the methods comprise administering to a cell an anti-breast cancer antibody that reduces or eliminates the biological activity of an endogenous breast cancer protein. Alternatively, the methods comprise administering to a cell or organism a recombinant nucleic acid encoding a breast cancer protein. This may be accomplished in any number of ways. In a preferred embodiment, e.g. when the breast cancer sequence is down-regulated in breast cancer, such state may be reversed by increasing the amount of breast cancer gene product in the cell. This can be accomplished, e.g., by overexpressing the endogenous breast cancer gene or administering a gene encoding the breast cancer sequence, using known gene-therapy techniques, e.g.. In a preferred embodiment, the gene therapy techniques include the incorporation of the exogenous gene using enhanced homologous recombination (EHR), e.g. as described in PCT/US93/03868, hereby incorporated by reference in its entirety. Alternatively, e.g. when the breast cancer sequence is up-regulated in breast cancer, the activity of the endogenous breast cancer gene is decreased, e.g. by the administration of a breast cancer antisense nucleic acid.

In one embodiment, the breast cancer proteins of the present invention may be used to generate polyclonal and monoclonal antibodies to breast cancer proteins. Similarly, the breast cancer proteins can be coupled, using standard technology, to affinity chromatography columns. These columns may then be used to purify breast cancer antibodies useful for production, diagnostic, or therapeutic purposes. In a preferred

embodiment, the antibodies are generated to epitopes unique to a breast cancer protein; that is, the antibodies show little or no cross-reactivity to other proteins. The breast cancer antibodies may be coupled to standard affinity chromatography columns and used to purify breast cancer proteins. The antibodies may also be used as blocking polypeptides, as outlined above, since they will specifically bind to the breast cancer protein.

Methods of identifying variant breast cancer-associated sequences

Without being bound by theory, expression of various breast cancer sequences is correlated with breast cancer. Accordingly, disorders based on mutant or variant breast cancer genes may be determined. In one embodiment, the invention provides methods for identifying cells containing variant breast cancer genes, e.g., determining all or part of the sequence of at least one endogenous breast cancer genes in a cell. This may be accomplished using any number of sequencing techniques. In a preferred embodiment, the invention provides methods of identifying the breast cancer genotype of an individual, e.g., determining all or part of the sequence of at least one breast cancer gene of the individual. This is generally done in at least one tissue of the individual, and may include the evaluation of a number of tissues or different samples of the same tissue. The method may include comparing the sequence of the sequenced breast cancer gene to a known breast cancer gene, i.e., a wild-type gene.

The sequence of all or part of the breast cancer gene can then be compared to the sequence of a known breast cancer gene to determine if any differences exist. This can be done using any number of known homology programs, such as Bestfit, etc. In a preferred embodiment, the presence of a difference in the sequence between the breast cancer gene of the patient and the known breast cancer gene correlates with a disease state or a propensity for a disease state, as outlined herein.

In a preferred embodiment, the breast cancer genes are used as probes to determine the number of copies of the breast cancer gene in the genome.

In another preferred embodiment, the breast cancer genes are used as probes to determine the chromosomal localization of the breast cancer genes. Information such as chromosomal localization finds use in providing a diagnosis or prognosis in particular when chromosomal abnormalities such as translocations, and the like are identified in the breast cancer gene locus.

Administration of pharmaceutical and vaccine compositions

In one embodiment, a therapeutically effective dose of a breast cancer protein or modulator thereof, is administered to a patient. By "therapeutically effective dose" herein is meant a dose that produces effects for which it is administered. The exact dose will depend on the purpose of the treatment, and will be ascertainable by one skilled in the art using known techniques (e.g., Ansel *et al.*, *Pharmaceutical Dosage Forms and Drug Delivery*; Lieberman, *Pharmaceutical Dosage Forms* (vols. 1-3, 1992), Dekker, ISBN 0824770846, 082476918X, 0824712692, 0824716981; Lloyd, *The Art, Science and Technology of Pharmaceutical Compounding* (1999); and Pickar, *Dosage Calculations* (1999)). As is known in the art, adjustments for breast cancer degradation, systemic versus localized delivery, and rate of new protease synthesis, as well as the age, body weight, general health, sex, diet, time of administration, drug interaction and the severity of the condition may be necessary, and will be ascertainable with routine experimentation by those skilled in the art. U.S. Patent Application N. 09/687,576, further discloses the use of compositions and methods of diagnosis and treatment in breast cancer is hereby expressly incorporated by reference.

A "patient" for the purposes of the present invention includes both humans and other animals, particularly mammals. Thus the methods are applicable to both human therapy and veterinary applications. In the preferred embodiment the patient is a mammal, preferably a primate, and in the most preferred embodiment the patient is human.

The administration of the breast cancer proteins and modulators thereof of the present invention can be done in a variety of ways as discussed above, including, but not limited to, orally, subcutaneously, intravenously, intranasally, transdermally, intraperitoneally, intramuscularly, intrapulmonary, vaginally, rectally, or intraocularly. In some instances, e.g., in the treatment of wounds and inflammation, the breast cancer proteins and modulators may be directly applied as a solution or spray.

The pharmaceutical compositions of the present invention comprise a breast cancer protein in a form suitable for administration to a patient. In the preferred embodiment, the pharmaceutical compositions are in a water soluble form, such as being present as pharmaceutically acceptable salts, which is meant to include both acid and base addition salts. "Pharmaceutically acceptable acid addition salt" refers to those salts that retain the

biological effectiveness of the free bases and that are not biologically or otherwise undesirable, formed with inorganic acids such as hydrochloric acid, hydrobromic acid, sulfuric acid, nitric acid, phosphoric acid and the like, and organic acids such as acetic acid, propionic acid, glycolic acid, pyruvic acid, oxalic acid, maleic acid, malonic acid, succinic acid, fumaric acid, tartaric acid, citric acid, benzoic acid, cinnamic acid, mandelic acid, methanesulfonic acid, ethanesulfonic acid, p-toluenesulfonic acid, salicylic acid and the like. "Pharmaceutically acceptable base addition salts" include those derived from inorganic bases such as sodium, potassium, lithium, ammonium, calcium, magnesium, iron, zinc, copper, manganese, aluminum salts and the like. Particularly preferred are the ammonium, potassium, sodium, calcium, and magnesium salts. Salts derived from pharmaceutically acceptable organic non-toxic bases include salts of primary, secondary, and tertiary amines, substituted amines including naturally occurring substituted amines, cyclic amines and basic ion exchange resins, such as isopropylamine, trimethylamine, diethylamine, triethylamine, tripropylamine, and ethanolamine.

The pharmaceutical compositions may also include one or more of the following: carrier proteins such as serum albumin; buffers; fillers such as microcrystalline cellulose, lactose, corn and other starches; binding agents; sweeteners and other flavoring agents; coloring agents; and polyethylene glycol.

The pharmaceutical compositions can be administered in a variety of unit dosage forms depending upon the method of administration. For example, unit dosage forms suitable for oral administration include, but are not limited to, powder, tablets, pills, capsules and lozenges. It is recognized that breast cancer protein modulators (e.g., antibodies, antisense constructs, ribozymes, small organic molecules, *etc.*) when administered orally, should be protected from digestion. This is typically accomplished either by complexing the molecule(s) with a composition to render it resistant to acidic and enzymatic hydrolysis, or by packaging the molecule(s) in an appropriately resistant carrier, such as a liposome or a protection barrier. Means of protecting agents from digestion are well known in the art.

The compositions for administration will commonly comprise a breast cancer protein modulator dissolved in a pharmaceutically acceptable carrier, preferably an aqueous carrier. A variety of aqueous carriers can be used, e.g., buffered saline and the like. These solutions are sterile and generally free of undesirable matter. These compositions may be sterilized by conventional, well known sterilization techniques. The compositions may

contain pharmaceutically acceptable auxiliary substances as required to approximate physiological conditions such as pH adjusting and buffering agents, toxicity adjusting agents and the like, e.g., sodium acetate, sodium chloride, potassium chloride, calcium chloride, sodium lactate and the like. The concentration of active agent in these formulations can vary widely, and will be selected primarily based on fluid volumes, viscosities, body weight and the like in accordance with the particular mode of administration selected and the patient's needs (e.g., *Remington's Pharmaceutical Science* (15th ed., 1980) and Goodman & Gillman, *The Pharmacologial Basis of Therapeutics* (Hardman *et al.*,eds., 1996)).

Thus, a typical pharmaceutical composition for intravenous administration would be about 0.1 to 10 mg per patient per day. Dosages from 0.1 up to about 100 mg per patient per day may be used, particularly when the drug is administered to a secluded site and not into the blood stream, such as into a body cavity or into a lumen of an organ. Substantially higher dosages are possible in topical administration. Actual methods for preparing parenterally administrable compositions will be known or apparent to those skilled in the art, *e.g.*, *Remington's Pharmaceutical Science* and Goodman and Gillman, *The Pharmacologial Basis of Therapeutics*, *supra*.

The compositions containing modulators of breast cancer proteins can be administered for therapeutic or prophylactic treatments. In therapeutic applications, compositions are administered to a patient suffering from a disease (e.g., a cancer) in an amount sufficient to cure or at least partially arrest the disease and its complications. An amount adequate to accomplish this is defined as a "therapeutically effective dose." Amounts effective for this use will depend upon the severity of the disease and the general state of the patient's health. Single or multiple administrations of the compositions may be administered depending on the dosage and frequency as required and tolerated by the patient. In any event, the composition should provide a sufficient quantity of the agents of this invention to effectively treat the patient. An amount of modulator that is capable of preventing or slowing the development of cancer in a mammal is referred to as a "prophylactically effective dose." The particular dose required for a prophylactic treatment will depend upon the medical condition and history of the mammal, the particular cancer being prevented, as well as other factors such as age, weight, gender, administration route, efficiency, *etc.* Such prophylactic treatments may be used, *e.g.*, in a mammal who has previously had cancer to prevent a

recurrence of the cancer, or in a mammal who is suspected of having a significant likelihood of developing cancer.

It will be appreciated that the present breast cancer protein-modulating compounds can be administered alone or in combination with additional breast cancer modulating compounds or with other therapeutic agent, *e.g.*, other anti-cancer agents or treatments.

In numerous embodiments, one or more nucleic acids, *e.g.*, polynucleotides comprising nucleic acid sequences set forth inTable 1, such as antisense polynucleotides or ribozymes, will be introduced into cells, *in vitro* or *in vivo*. The present invention provides methods, reagents, vectors, and cells useful for expression of breast cancer-associated polypeptides and nucleic acids using *in vitro* (cell-free), *ex vivo* or *in vivo* (cell or organism-based) recombinant expression systems.

The particular procedure used to introduce the nucleic acids into a host cell for expression of a protein or nucleic acid is application specific. Many procedures for introducing foreign nucleotide sequences into host cells may be used. These include the use of calcium phosphate transfection, spheroplasts, electroporation, liposomes, microinjection, plasma vectors, viral vectors and any of the other well known methods for introducing cloned genomic DNA, cDNA, synthetic DNA or other foreign genetic material into a host cell (*see, e.g.,* Berger & Kimmel, *Guide to Molecular Cloning Techniques, Methods in Enzymology* volume 152 (Berger), Ausubel *et al.*, eds., *Current Protocols* (supplemented through 1999), and Sambrook *et al.*, *Molecular Cloning - A Laboratory Manual* (2nd ed., Vol. 1-3, 1989.

In a preferred embodiment, breast cancer proteins and modulators are administered as therapeutic agents, and can be formulated as outlined above. Similarly, breast cancer genes (including both the full-length sequence, partial sequences, or regulatory sequences of the breast cancer coding regions) can be administered in a gene therapy application. These breast cancer genes can include antisense applications, either as gene therapy (i.e. for incorporation into the genome) or as antisense compositions, as will be appreciated by those in the art.

Breast cancer polypeptides and polynucleotides can also be administered as vaccine compositions to stimulate HTL, CTL and antibody responses.. Such vaccine compositions can include, e.g., lipidated peptides (*see, e.g.,*Vitiello, A. *et al.*, *J. Clin. Invest.* 95:341 (1995)), peptide compositions encapsulated in poly(DL-lactide-co-glycolide) ("PLG")

microspheres (*see, e.g.,* Eldridge, *et al.*, *Molec. Immunol.* 28:287-294, (1991); Alonso *et al.*, *Vaccine* 12:299-306 (1994); Jones *et al.*, *Vaccine* 13:675-681 (1995)), peptide compositions contained in immune stimulating complexes (ISCOMS) (*see, e.g.,* Takahashi *et al.*, *Nature* 344:873-875 (1990); Hu *et al.*, *Clin Exp Immunol.* 113:235-243 (1998)), multiple antigen peptide systems (MAPs) (*see, e.g.*, Tam, *Proc. Natl. Acad. Sci. U.S.A.* 85:5409-5413 (1988); Tam, *J. Immunol. Methods* 196:17-32 (1996)), peptides formulated as multivalent peptides; peptides for use in ballistic delivery systems, typically crystallized peptides, viral delivery vectors (Perkus, *et al.*, In: *Concepts in vaccine development* (Kaufmann, ed., p. 379, 1996); Chakrabarti, *et al.*, *Nature* 320:535 (1986); Hu *et al.*, *Nature* 320:537 (1986); Kieny, *et al.*, *AIDS Bio/Technology* 4:790 (1986); Top *et al.*, *J. Infect. Dis.* 124:148 (1971); Chanda *et al.*, *Virology* 175:535 (1990)), particles of viral or synthetic origin (*see, e.g.*, Kofler *et al.*, *J. Immunol. Methods.* 192:25 (1996); Eldridge *et al.*, *Sem. Hematol.* 30:16 (1993); Falo *et al.*, *Nature Med.* 7:649 (1995)), adjuvants (Warren *et al.*, *Annu. Rev. Immunol.* 4:369 (1986); Gupta *et al.*, *Vaccine* 11:293 (1993)), liposomes (Reddy *et al.*, *J. Immunol.* 148:1585 (1992); Rock, *Immunol. Today* 17:131 (1996)), or, naked or particle absorbed cDNA (Ulmer, *et al.*, *Science* 259:1745 (1993); Robinson *et al.*, *Vaccine* 11:957 (1993); Shiver *et al.*, In: *Concepts in vaccine development* (Kaufmann, ed., p. 423, 1996); Cease & Berzofsky, *Annu. Rev. Immunol.* 12:923 (1994) and Eldridge *et al.*, *Sem. Hematol.* 30:16 (1993)). Toxin-targeted delivery technologies, also known as receptor mediated targeting, such as those of Avant Immunotherapeutics, Inc. (Needham, Massachusetts) may also be used.

Vaccine compositions often include adjuvants. Many adjuvants contain a substance designed to protect the antigen from rapid catabolism, such as aluminum hydroxide or mineral oil, and a stimulator of immune responses, such as lipid A, *Bortadella pertussis* or *Mycobacterium tuberculosis* derived proteins. Certain adjuvants are commercially available as, e.g., Freund's Incomplete Adjuvant and Complete Adjuvant (Difco Laboratories, Detroit, MI); Merck Adjuvant 65 (Merck and Company, Inc., Rahway, NJ); AS-2 (SmithKline Beecham, Philadelphia, PA); aluminum salts such as aluminum hydroxide gel (alum) or aluminum phosphate; salts of calcium, iron or zinc; an insoluble suspension of acylated tyrosine; acylated sugars; cationically or anionically derivatized polysaccharides; polyphosphazenes; biodegradable microspheres; monophosphoryl lipid A and quil A. Cytokines, such as GM-CSF, interleukin-2, -7, -12, and other like growth factors, may also be used as adjuvants.

Vaccines can be administered as nucleic acid compositions wherein DNA or RNA encoding one or more of the polypeptides, or a fragment thereof, is administered to a patient. This approach is described, for instance, in Wolff *et. al.*, *Science* 247:1465 (1990) as well as U.S. Patent Nos. 5,580,859; 5,589,466; 5,804,566; 5,739,118; 5,736,524; 5,679,647; WO 98/04720; and in more detail below. Examples of DNA-based delivery technologies include "naked DNA", facilitated (bupivicaine, polymers, peptide-mediated) delivery, cationic lipid complexes, and particle-mediated ("gene gun") or pressure-mediated delivery (*see, e.g.,* U.S. Patent No. 5,922,687).

For therapeutic or prophylactic immunization purposes, the peptides of the invention can be expressed by viral or bacterial vectors. Examples of expression vectors include attenuated viral hosts, such as vaccinia or fowlpox. This approach involves the use of vaccinia virus, e.g., as a vector to express nucleotide sequences that encode breast cancer polypeptides or polypeptide fragments. Upon introduction into a host, the recombinant vaccinia virus expresses the immunogenic peptide, and thereby elicits an immune response. Vaccinia vectors and methods useful in immunization protocols are described in, *e.g.*, U.S. Patent No. 4,722,848. Another vector is BCG (Bacille Calmette Guerin). BCG vectors are described in Stover *et al.*, *Nature* 351:456-460 (1991). A wide variety of other vectors useful for therapeutic administration or immunization e.g. adeno and adeno-associated virus vectors, retroviral vectors, *Salmonella typhi* vectors, detoxified anthrax toxin vectors, and the like, will be apparent to those skilled in the art from the description herein (*see, e.g.,* Shata *et al.*, *Mol Med Today* 6:66-71 (2000); Shedlock *et al.*, *J Leukoc Biol* 68:793-806 (2000); Hipp *et al.*, *In Vivo* 14:571-85 (2000)).

Methods for the use of genes as DNA vaccines are well known, and include placing a breast cancer gene or portion of a breast cancer gene under the control of a regulatable promoter or a tissue-specific promoter for expression in a breast cancer patient. The breast cancer gene used for DNA vaccines can encode full-length breast cancer proteins, but more preferably encodes portions of the breast cancer proteins including peptides derived from the breast cancer protein. In one embodiment, a patient is immunized with a DNA vaccine comprising a plurality of nucleotide sequences derived from a breast cancer gene. For example, breast cancer-associated genes or sequence encoding subfragments of a breast cancer protein are introduced into expression vectors and tested for their immunogenicity in the context of Class I MHC and an ability to generate cytotoxic T cell responses. This

procedure provides for production of cytotoxic T cell responses against cells which present antigen, including intracellular epitopes.

In a preferred embodiment, the DNA vaccines include a gene encoding an adjuvant molecule with the DNA vaccine. Such adjuvant molecules include cytokines that increase the immunogenic response to the breast cancer polypeptide encoded by the DNA vaccine. Additional or alternative adjuvants are available.

In another preferred embodiment breast cancer genes find use in generating animal models of breast cancer. When the breast cancer gene identified is repressed or diminished in cancer tissue, gene therapy technology, e.g., wherein antisense RNA directed to the breast cancer gene will also diminish or repress expression of the gene. Animal models of breast cancer find use in screening for modulators of a breast cancer-associated sequence or modulators of breast cancer. Similarly, transgenic animal technology including gene knockout technology, e.g. as a result of homologous recombination with an appropriate gene targeting vector, will result in the absence or increased expression of the breast cancer protein. When desired, tissue-specific expression or knockout of the breast cancer protein may be necessary.

It is also possible that the breast cancer protein is overexpressed in breast cancer. As such, transgenic animals can be generated that overexpress the breast cancer protein. Depending on the desired expression level, promoters of various strengths can be employed to express the transgene. Also, the number of copies of the integrated transgene can be determined and compared for a determination of the expression level of the transgene. Animals generated by such methods find use as animal models of breast cancer and are additionally useful in screening for modulators to treat breast cancer.

Kits for Use in Diagnostic and/or Prognostic Applications

For use in diagnostic, research, and therapeutic applications suggested above, kits are also provided by the invention. In the diagnostic and research applications such kits may include any or all of the following: assay reagents, buffers, breast cancer-specific nucleic acids or antibodies, hybridization probes and/or primers, antisense polynucleotides, ribozymes, dominant negative breast cancer polypeptides or polynucleotides, small molecules inhibitors of breast cancer-associated sequences *etc.* A therapeutic product may include sterile saline or another pharmaceutically acceptable emulsion and suspension base.

In addition, the kits may include instructional materials containing directions (i.e., protocols) for the practice of the methods of this invention. While the instructional materials typically comprise written or printed materials they are not limited to such. Any medium capable of storing such instructions and communicating them to an end user is contemplated by this invention. Such media include, but are not limited to electronic storage media (e.g., magnetic discs, tapes, cartridges, chips), optical media (e.g., CD ROM), and the like. Such media may include addresses to internet sites that provide such instructional materials.

The present invention also provides for kits for screening for modulators of breast cancer-associated sequences. Such kits can be prepared from readily available materials and reagents. For example, such kits can comprise one or more of the following materials: a breast cancer-associated polypeptide or polynucleotide, reaction tubes, and instructions for testing breast cancer-associated activity. Optionally, the kit contains biologically active breast cancer protein. A wide variety of kits and components can be prepared according to the present invention, depending upon the intended user of the kit and the particular needs of the user. Diagnosis would typically involve evaluation of a plurality of genes or products. The genes will be selected based on correlations with important parameters in disease which may be identified in historical or outcome data.

EXAMPLES

Example 1: Tissue Preparation, Labeling Chips, and Fingerprints

Purifying total RNA from tissue sample using TRIzol Reagent

The sample weight is first estimated. The tissue samples are homogenized in 1 ml of TRIzol per 50 mg of tissue using a homogenizer (e.g., Polytron 3100). The size of the generator/probe used depends upon the sample amount. A generator that is too large for the amount of tissue to be homogenized will cause a loss of sample and lower RNA yield. A larger generator (e.g., 20 mm) is suitable for tissue samples weighing more than 0.6 g. Fill tubes should not be overfilled. If the working volume is greater than 2 ml and no greater than 10 ml, a 15 ml polypropylene tube (Falcon 2059) is suitable for homogenization.

Tissues should be kept frozen until homogenized. The TRIzol is added directly to the frozen tissue before homogenization. Following homogenization, the insoluble material is removed from the homogenate by centrifugation at 7500 x g for 15 min. in a Sorvall superspeed or 12,000 x g for 10 min. in an Eppendorf centrifuge at 4°C. The cleared homogenate is then transferred to a new tube(s). Samples may be frozen and stored at -60 to -70°C for at least one month or else continue with the purification.

The next process is phase separation. The homogenized samples are incubated for 5 minutes at room temperature. Then, 0.2 ml of chloroform per 1ml of TRIzol reagent is added to the homogenization mixture. The tubes are securely capped and shaken vigorously by hand (do not vortex) for 15 seconds. The samples are then incubated at room temp. for 2-3 minutes and next centrifuged at 6500 rpm in a Sorvall superspeed for 30 min. at 4oC.

The next process is RNA Precipitation. The aqueous phase is transferred to a fresh tube. The organic phase can be saved if isolation of DNA or protein is desired. Then 0.5 ml of isopropyl alcohol is added per 1ml of TRIzol reagent used in the original homogenization. Then, the tubes are securely capped and inverted to mix. The samples are then incubated at room temp. for 10 minutes an centrifuged at 6500 rpm in Sorvall for 20 min. at 4°C.

The RNA is then washed. The supernatant is poured off and the pellet washed with cold 75% ethanol. 1 ml of 75% ethanol is used per 1 ml of the TRIzol reagent used in the initial homogenization. The tubes are capped securely and inverted several times to loosen pellet without vortexing . They are next centrifuged at <8000 rpm (<7500 x g) for 5 minutes at 4°C.

The RNA wash is decanted. The pellet is carefully transferred to an Eppendorf tube (sliding down the tube into the new tube by use of a pipet tip to help guide it in if necessary). Tube(s) sizes for precipitating the RNA depending on the working volumes. Larger tubes may take too long to dry. Dry pellet. The RNA is then resuspended in an appropriate volume (e.g., 2 -5 ug/ul) of DEPC H_20. The absorbance is then measured.

The poly A+ mRNA may next be purified from total RNA by other methods such as Qiagen' s RNeasy kit. The poly A $^+$ mRNA is purified from total RNA by adding the oligotex suspension which has been heated to 37°C and mixing prior to adding to RNA. The Elution Buffer is incubated at 70°C. If there is precipitate in the buffer, warm up the 2 x

Binding Buffer at 65°C. The the total RNA is mixed with DEPC-treated water, 2 x Binding Buffer, and Oligotex according to Table 2 on page 16 of the Oligotex Handbook and next incubated for 3 minutes at 65°C and 10 minutes at room temperature.

The preparation is centrifuged for 2 minutes at 14,000 to 18,000 g, preferably, at a "soft setting," The supernatant is removed without disturbing Oligotex pellet. A little bit of solution can be left behind to reduce the loss of Oligotex. The supernatant is saved until satisfactory binding and elution of poly A^+ mRNA has been found.

Then, the preparation is gently resuspended in Wash Buffer OW2 and pipetted onto the spin column and centrifuged at full speed (soft setting if possible) for 1 minute.

Next, the spin column is transferred to a new collection tube and gently resuspended in Wash Buffer OW2 and centrifuged as described herein.

Then, the spin column is transferred to a new tube and eluted with 20 to 100 ul of preheated (70°C) Elution Buffer. The Oligotex resin is gently resuspended by pipetting up and down. The centrifugation is repeated as above and the elution repeated with fresh elution buffer or first eluate to keep the elution volume low.

The absorbance is next read to determine the yield, using diluted Elution Buffer as the blank.

Before proceeding with cDNA synthesis, the mRNA is precipitated before proceeding with cDNA synthesis, as components leftover or in the Elution Buffer from the Oligotex purification procedure will inhibit downstream enzymatic reactions of the mRNA. 0.4 vol. of 7.5 M NH4OAc + 2.5 vol. of cold 100% ethanol is added and the preparation precipitated at -20°C 1 hour to overnight (or 20-30 min. at -70°C), and centrifuged at 14,000-16,000 x g for 30 minutes at 4°C. Next, the pellet is washed with 0.5 ml of 80% ethanol (-20°C) and then centrifuged at 14,000-16,000 x g for 5 minutes at room temperature. The80% ethanol wash is then repeated. The last bit of ethanol from the pellet is then dried without use of a speed vacuum and the pellet is then resuspended in DEPC H_20 at 1ug/ul concentration.

Alternatively the RNA may be purified using other methods (e.g., Qiagen's RNeasy kit).

No more than 100 ug is added to the RNeasy column. The sample volume is adjusted to 100 ul with RNase-free water. 350 ul Buffer RLT and then 250 ul ethanol

(100%) are added to the sample. The preparation is then mixed by pipetting and applied to an RNeasy mini spin column for centrifugation (15 sec at >10,000 rpm). If yield is low, reapply the flowthrough to the column and centrifuge again.

Then, transfer column to a new 2 ml collection tube and add 500 ul Buffer RPE and centrifuge for 15 sec at >10,000 rpm. The flowthrough is discarded. 500 ul Buffer RPE and is then added and the preparation is centriuged for 15 sec at >10,000 rpm. The flowthrough is discarded. and the column membrane dried by centrifuging for 2 min at maximum speed. The column is transferred to a new 1.5-ml collection tube. 30-50 ul of RNase-free water is applied directly onto column membrane. The column is then centrifuged for 1 min at >10,000 rpm and the elution step repeated.

The absorbance is then read to determine yield. If necessary, the material may be ethanol precipitated with ammonium acetate and 2.5X volume 100% ethanol.

First Strand cDNA Synthesis

The first strand can be make using using Gibco's "SuperScript Choice System for cDNA Synthesis" kit. The starting material is 5 ug of total RNA or 1 ug of polyA+ mRNAl. For total RNA, 2 ul of SuperScript RT is used; for polyA+ mRNA, 1 ul of SuperScript RT is used. The final volume of first strand synthesis mix is 20 ul. The RNA should be in a volume no greater than 10 ul. The RNA is incubated with 1 ul of 100 pmol T7-T24 oligo for 10 min at 70°C followed by addition on ice of 7 ul of: 4ul 5X 1st Strand Buffer, 2 ul of 0.1M DTT, and 1 ul of 10mM dNTP mix. The preparation is then incubated at 37°C for 2 min before addition of the SuperScript RT followed by incubation at 37°C for 1 hour.

Second Strand Synthesis

For the second strand synthesis, place 1st strand reactions on ice and add: 91 ul DEPC H_20; 30 ul 5X 2nd Strand Buffer; 3 ul 10mM dNTP mix; 1 ul 10 U/ul E.coli DNA Ligase; 4 ul 10 U/ul E.coli DNA Polymerase; and 1 ul 2 U/ul RNase H. Mix and incubate 2 hours at 16°C. Add 2 ul T4 DNA Polymerase. Incubate 5 min at 16°C. Add 10 ul of 0.5M EDTA.

Cleaning up cDNA

The cDNA is purified using Phenol:Chloroform:Isoamyl Alcohol (25:24:1) and Phase-Lock gel tubes. The PLG tubes are centrifuged for 30 sec at maximum speed. The cDNA mix is then transferred to PLG tube. An equal volume of phenol:chloroform:isamyl alcohol is then added, the preparation shaken vigorously (no vortexing), and centrifuged for 5 minutes at maximum speed. The top aqueous solution is transferred to a new tube and ethanol precipitated by adding 7.5X 5M NH4OAc and 2.5X volume of 100% ethanol. Next, it is centrifuged immediately at room temperature for 20 min, maximum speed. The supernatant is removed, and the pellet washed with 2X with cold 80% ethanol. As much ethanol wash as possible should be removed before air drying the pellet; and resuspending it in 3 ul RNase-free water.

In vitro Transcription (IVT) and labeling with biotin

In vitro Transcription (IVT) and labeling with biotin is performed as follows: Pipet 1.5 ul of cDNA into a thin-wall PCR tube. Make NTP labeling mix by combining 2 ul T7 10xATP (75 mM) (Ambion); 2 ul T7 10xGTP (75 mM) (Ambion); 1.5 ul T7 10xCTP (75 mM) (Ambion); 1.5 ul T7 10xUTP (75 mM) (Ambion); 3.75 ul 10 mM Bio-11-UTP (Boehringer-Mannheim/Roche or Enzo); 3.75 ul 10 mM Bio-16-CTP (Enzo); 2 ul 10x T7 transcription buffer (Ambion); and 2 ul 10x T7 enzyme mix (Ambion). The final volume is 20 ul. Incubate 6 hours at 37°C in a PCR machine. The RNA can be furthered cleaned. Clean-up follows the previous instructions for RNeasy columns or Qiagen's RNeasy protocol handbook. The cRNA often needs to be ethanol precipitated by resuspension in a volume compatible with the fragmentation step.

Fragmentation is performed as follows. 15 ug of labeled RNA is usually fragmented. Try to minimize the fragmentation reaction volume; a 10 ul volume is recommended but 20 ul is all right. Do not go higher than 20 ul because the magnesium in the fragmentation buffer contributes to precipitation in the hybridization buffer. Fragment RNA by incubation at 94 C for 35 minutes in 1 x Fragmentation buffer (5 x Fragmentation buffer is 200 mM Tris-acetate, pH 8.1; 500 mM KOAc; 150 mM MgOAc). The labeled RNA transcript can be analyzed before and after fragmentation. Samples can be heated to 65°C for 15 minutes and electrophoresed on 1% agarose/TBE gels to get an approximate idea of the transcript size range

For hybridization, 200 ul (10 ug cRNA) of a hybridization mix is put on the chip. If multiple hybridizations are to be done (such as cycling through a 5 chip set), then it is recommended that an initial hybridization mix of 300 ul or more be made. The hybridization mix is: fragment labeled RNA (50 ng/ul final conc.); 50 pM 948-b control oligo; 1.5 pM BioB; 5 pM BioC; 25 pM BioD; 100 pM CRE; 0.1 mg/ml herring sperm DNA; 0.5 mg/ml acetylated BSA; and 300 ul with 1xMES hyb buffer.

The hybridization reaction is conducted with non-biotinylated IVT (purified by RNeasy columns) (see example 1 for steps from tissue to IVT): The following mixture is prepared:

IVT antisense RNA; 4 μg:	μl
Random Hexamers (1 μg/μl):	4 μl
H_2O:	μl
	14 μl

Incubate the above 14 μl mixture at 70°C for 10 min.; then put on ice.

The Reverse transcription procedure uses the following mixture:

0.1 M DTT:	3 μl
50X dNTP mix:	0.6 μl
H_2O:	2.4 μl
Cy3 or Cy5 dUTP (1mM):	3 μl
SS RT II (BRL):	1 μl
	16 μl

The above solution is added to the hybridization reaction and incubated for 30 min., 42°C. Then, 1 μl SSII is added and incubated for another hour before being placed on ice.

The 50X dNTP mix contains 25mM of cold dATP, dCTP, and dGTP, 10mM of dTTP and is made by adding 25 μl each of 100mM dATP, dCTP, and dGTP; 10 μl of 100mM dTTP to 15 μl H_2O.]

RNA degradation is performed as follows. Add 86 μl H2O, 1.5 μl 1M NaOH/ 2 mM EDTA and incubate at 65°C, 10 min.. For U-Con 30, 500 μl TE/sample spin at 7000 g for 10 min, save flow through for purification. For Qiagen purification, suspend u-con

recovered material in 500 µl buffer PB and proceed using Qiagen protocol. For DNAse digestion, add 1 ul of 1/100 dilution of DNAse/30 ul Rx and incubate at 37°C for 15 min. Incubate at 5 min 95°C to denature the DNAse.

Sample preparation

For sample preparation, add Cot-1 DNA, 10 l; 50X dNTPs, 1 l; 20X SSC, 2.3 l; Na pyro phosphate, 7.5 l; 10 mg/ml Herring sperm DNA; 1 ul of 1/10 dilution to 21.8 final vol. Dry in speed vac. Resuspend in 15 l H_20. Add 0.38 l 10% SDS. Heat 95°C, 2 min and slow cool at room temp. for 20 min. Put on slide and hybridize overnight at 64°C. Washing after the hybridization: 3X SSC/0.03% SDS: 2 min., 37.5 mls 20X SSC+0.75mls 10% SDS in 250mls H2O; 1X SSC: 5 min., 12.5 mls 20X SSC in 250mls H2O; 0.2X SSC: 5 min., 2.5 mls 20X SSC in 250mls H2O. Dry slides and scan at appropiate PMT's and channels.

Table 1

BCA4 DNA sequence
Gene name: osteoblast specific factor 2 (periostin); Unigene number: Hs.136348; Probeset Accession #: D13666; Nucleic Acid Accession #: NM_006475; Coding sequence: 12-2522 (start and stop codons underlined)

```
AGAGACTCAA GATGATTCCC TTTTTACCCA TGTTTTCTCT ACTATTGCTG CTTATTGTTA    60
ACCCTATAAA CGCCAACAAT CATTATGACA AGATCTTGGC TCATAGTCGT ATCAGGGGTC   120
GGGACCAAGG CCCAAATGTC TGTGCCCTTC AACAGATTTT GGGCACCAAA AAGAAATACT   180
TCAGCACTTG TAAGAACTGG TATAAAAAGT CCATCTGTGG ACAGAAAACG ACTGTTTTAT   240
ATGAATGTTG CCCTGGTTAT ATGAGAATGG AAGGAATGAA AGGCTGCCCA GCAGTTTTGC   300
CCATTGACCA TGTTTATGGC ACTCTGGGCA TCGTGGGAGC CACCACAACG CAGCGCTATT   360
CTGACGCCTC AAAACTGAGG GAGGAGATCG AGGGAAAGGG ATCCTTCACT TACTTTGCAC   420
CGAGTAATGA GGCTTGGGAC AACTTGGATT CTGATATCCG TAGAGGTTTG GAGAGCAACG   480
TGAATGTTGA ATTACTGAAT GCTTTACATA GTCACATGAT TAATAAGAGA ATGTTGACCA   540
AGGACTTAAA AAATGGCATG ATTATTCCTT CAATGTATAA CAATTTGGGG CTTTTCATTA   600
ACCATTATCC TAATGGGGTT GTCACTGTTA ATTGTGCTCG AATCATCCAT GGGAACCAGA   660
TTGCAACAAA TGGTGTTGTC CATGTCATTG ACCGTGTGCT TACACAAATT GGTACCTCAA   720
TTCAAGACTT CATTGAAGCA GAAGATGACC TTTCATCTTT TAGAGCAGCT GCCATCACAT   780
CGGACATATT GGAGGCCCTT GGAAGAGACG GTCACTTCAC ACTCTTTGCT CCCACCAATG   840
AGGCTTTTGA GAAACTTCCA CGAGGTGTCC TAGAAAGGTT CATGGGAGAC AAAGTGGCTT   900
CCGAAGCTCT TATGAAGTAC CACATCTTAA ATACTCTCCA GTGTTCTGAG TCTATTATGG   960
GAGGAGCAGT CTTTGAGACG CTGGAAGGAA ATACAATTGA GATAGGATGT GACGGTGACA  1020
GTATAACAGT AAATGGAATC AAAATGGTGA ACAAAAAGGA TATTGTGACA AATAATGGTG  1080
TGATCCATTT GATTGATCAG GTCCTAATTC CTGATTCTGC CAAACAAGTT ATTGAGCTGG  1140
CTGGAAAACA GCAAACCACC TTCACGGATC TTGTGGCCCA ATTAGGCTTG GCATCTGCTC  1200
TGAGGCCAGA TGGAGAATAC ACTTTGCTGG CACCTGTGAA TAATGCATTT TCTGATGATA  1260
CTCTCAGCAT GGTTCAGCGC CTCCTTAAAT TAATTCTGCA GAATCACATA TTGAAAGTAA  1320
AAGTTGGCCT TAATGAGCTT TACAACGGGC AAATACTGGA AACCATCGGA GGCAAACAGC  1380
TCAGAGTCTT CGTATATCGT ACAGCTGTCT GCATTGAAAA TTCATGCATG GAGAAAGGGA  1440
GTAAGCAAGG GAGAAACGGT GCGATTCACA TATTCCGCGA GATCATCAAG CCAGCAGAGA  1500
AATCCCTCCA TGAAAAGTTA AAACAAGATA AGCGCTTTAG CACCTTCCTC AGCCTACTTG  1560
AAGCTGCAGA CTTGAAAGAG CTCCTGACAC AACCTGGAGA CTGGACATTA TTTGTGCCAA  1620
CCAATGATGC TTTTAAGGGA ATGACTAGTG AAGAAAAAGA AATTCTGATA CGGGACAAAA  1680
ATGCTCTTCA AAACATCATT CTTTATCACC TGACACCAGG AGTTTTCATT GGAAAAGGAT  1740
TTGAACCTGG TGTTACTAAC ATTTTAAAGA CCACACAAGG AAGCAAAATC TTTCTGAAAG  1800
AAGTAAATGA TACACTTCTG GTGAATGAAT TGAAATCAAA AGAATCTGAC ATCATGACAA  1860
CAAATGGTGT AATTCATGTT GTAGATAAAC TCCTCTATCC AGCAGACACA CCTGTTGGAA  1920
ATGATCAACT GCTGGAAATA CTTAATAAAT TAATCAAATA CATCCAAATT AAGTTTGTTC  1980
GTGGTAGCAC CTTCAAAGAA ATCCCCGTGA CTGTCTATAC AACTAAAATT ATAACCAAAG  2040
TTGTGGAACC AAAAATTAAA GTGATTGAAG GCAGTCTTCA GCCTATTATC AAAACTGAAG  2100
GACCCACACT AACAAAAGTC AAAATTGAAG GTGAACCTGA ATTCAGACTG ATTAAAGAAG  2160
GTGAAACAAT AACTGAAGTG ATCCATGGAG AGCCAATTAT TAAAAAATAC ACCAAAATCA  2220
TTGATGGAGT GCCTGTGGAA ATAACTGAAA AAGAGACACG AGAAGAACGA ATCATTACAG  2280
GTCCTGAAAT AAAATACACT AGGATTTCTA CTGGAGGTGG AGAAACAGAA GAAACTCTGA  2340
AGAAATTGTT ACAAGAAGAG GTCACCAAGG TCACCAAATT CATTGAAGGT GGTGATGGTC  2400
ATTTATTTGA AGATGAAGAA ATTAAAAGAC TGCTTCAGGG AGACACACCC GTGAGGAAGT  2460
TGCAAGCCAA CAAAAAAGTT CAAGGTTCTA GAAGACGATT AAGGGAAGGT CGTTCTCAGT  2520
GAAAATCCAA AAACCAGAAA AAAATGTTTA TACAACCCTA AGTCAATAAC CTGACCTTAG  2580
AAAATTGTGA GAGCCAAGTT GACTTCAGGA ACTGAAACAT CAGCACAAAG AAGCAATCAT  2640
CAAATAATTC TGAACACAAA TTTAATATTT TTTTTTCTGA ATGAGAAACA TGAGGGAAAT  2700
TGTGGAGTTA GCCTCCTGTG GTAAAGGAAT TGAAGAAAAT ATAACACCTT ACACCCTTTT  2760
TCATCTTGAC ATTAAAAGTT CTGGCTAACT TTGGAATCCA TTAGAGAAAA ATCCTTGTCA  2820
CCAGATTCAT TACAATTCAA ATCGAAGAGT TGTGAACTGT TATCCCATTG AAAAGACCGA  2880
GCCTTGTATG TATGTTATGG ATACATAAAA TGCACGCAAG CCATTATCTC TCCATGGGAA  2940
GCTAAGTTAT AAAAATAGGT GCTTGGTGTA CAAAACTTTT TATATCAAAA GGCTTTGCAC  3000
ATTTCTATAT GAGTGGGTTT ACTGGTAAAT TATGTTATTT TTTACAACTA ATTTTGTACT  3060
CTCAGAATGT TTGTCATATG CTTCTTGCAA TGCATATTTT TTAATCTCAA ACGTTTCAAT  3120
AAAACCATTT TTCAGATATA AAGAGAATTA CTTCAAATTG AGTAATTCAG AAAAACTCAA  3180
GATTTAAGTT AAAAAGTGGT TTGGACTTGG GAA
```

BCA4 Protein sequence
Gene name: osteoblast specific factor 2 (periostin); Unigene number: Hs.136348; Probeset Accession #: D13666; Protein Accession #: NP_006466; Predicted Signal sequence: 1-21; TM domains: none; PFAM domains: fasciclin_domains: 94-232, 234-367, 496-630; Summary: a secreted protein involved in adhesion and osteoblast development; may participate in preferential metastasis of breast cancer to bone.

```
MIPFLPMFSL LLLLIVNPIN ANNHYDKILA HSRIRGRDQG PNVCALQQIL GTKKKYFSTC    60
KNWYKKSICG QKTTVLYECC PGYMRMEGMK GCPAVLPIDH VYGTLGIVGA TTTQRYSDAS   120
KLREEIEGKG SFTYFAPSNE AWDNLDSDIR RGLESNVNVE LLNALHSHMI NKRMLTKDLK   180
```

NGMIIPSMYN NLGLFINHYP NGVVTVNCAR IIHGNQIATN GVVHVIDRVL TQIGTSIQDF 240
IEAEDDLSSF RAAAITSDIL EALGRDGHFT LFAPTNEAFE KLPRGVLERF MGDKVASEAL 300
MKYHILNTLQ CSESIMGGAV FETLEGNTIE IGCDGDSITV NGIKMVNKKD IVTNNGVIHL 360
IDQVLIPDSA KQVIELAGKQ QTTFTDLVAQ LGLASALRPD GEYTLLAPVN NAFSDDTLSM 420
VQRLLKLILQ NHILKVKVGL NELYNGQILE TIGGKQLRVF VYRTAVCIEN SCMEKGSKQG 480
RNGAIHIFRE IIKPAEKSLH EKLKQDKRFS TFLSLLEAAD LKELLTQPGD WTLFVPTNDA 540
FKGMTSEEKE ILIRDKNALQ NIILYHLTPG VFIGKGFEPG VTNILKTTQG SKIFLKEVND 600
TLLVNELKSK ESDIMTTNGV IHVVDKLLYP ADTPVGNDQL LEILNKLIKY IQIKFVRGST 660
FKEIPVTVYT TKIITKVVEP KIKVIEGSLQ PIIKTEGPTL TKVKIEGEPE FRLIKEGETI 720
TEVIHGEPII KKYTKIIDGV PVEITEKETR EERIITGPEI KYTRISTGGG ETEETLKKLL 780
QEEVTKVTKF IEGGDGHLFE DEEIKRLLQG DTPVRKLQAN KKVQGSRRRL REGRSQ

BCA7 DNA sequence
Gene name: 5T4 oncofetal trophoblast glycoprotein; Unigene number: Hs.82128; Probeset Accession #: Z29083; Nucleic Acid Accession #: NM_006670; Coding sequence: 85-1347 (start and stop codons underlined)

CCGGCTCGCG CCCTCCGGGC CCAGCCTCCC GAGCCTTCGG AGCGGGCGCC GTCCCAGCCC 60
AGCTCCGGGG AAACGCGAGC CGCGATGCCT GGGGGGTGCT CCCGGGGCCC CGCCGCCGGG 120
GACGGGCGTC TGCGGCTGGC GCGACTAGCG CTGGTACTCC TGGGCTGGGT CTCCTCGTCT 180
TCTCCCACCT CCTCGGCATC CTCCTTCTCC TCCTCGGCGC CGTTCCTGGC TTCCGCCGTG 240
TCCGCCCAGC CCCCGCTGCC GGACCAGTGC CCCGCGCTGT GCGAGTGCTC CGAGGCAGCG 300
CGCACAGTCA AGTGCGTTAA CCGCAATCTG ACCGAGGTGC CCACGGACCT GCCCGCCTAC 360
GTGCGCAACC TCTTCCTTAC CGGCAACCAG CTGGCCGTGC TCCCTGCCGG CGCCTTCGCC 420
CGCCGGCCGC CGCTGGCGGA GCTGGCCGCG CTCAACCTCA GCGGCAGCCG CCTGGACGAG 480
GTGCGCGCGG GCGCCTTCGA GCATCTGCCC AGCCTGCGCC AGCTCGACCT CAGCCACAAC 540
CCACTGGCCG ACCTCAGTCC CTTCGCTTTC TCGGGCAGCA ATGCCAGCGT CTCGGCCCCC 600
AGTCCCCTTG TGGAACTGAT CCTGAACCAC ATCGTGCCCC CTGAAGATGA GCGGCAGAAC 660
CGGAGCTTCG AGGGCATGGT GGTGGCGGCC CTGCTGGCGG GCCGTGCACT GCAGGGGCTC 720
CGCCGCTTGG AGCTGGCCAG CAACCACTTC CTTTACCTGC CGCGGGATGT GCTGGCCCAA 780
CTGCCCAGCC TCAGGCACCT GGACTTAAGT AATAATTCGC TGGTGAGCCT GACCTACGTG 840
TCCTTCCGCA ACCTGACACA TCTAGAAAGC CTCCACCTGG AGGACAATGC CCTCAAGGTC 900
CTTCACAATG GCACCCTGGC TGAGTTGCAA GGTCTACCCC ACATTAGGGT TTTCCTGGAC 960
AACAATCCCT GGGTCTGCGA CTGCCACATG GCAGACATGG TGACCTGGCT CAAGGAAACA 1020
GAGGTAGTGC AGGGCAAAGA CCGGCTCACC TGTGCATATC CGGAAAAAAT GAGGAATCGG 1080
GTCCTCTTGG AACTCAACAG TGCTGACCTG GACTGTGACC CGATTCTTCC CCCATCCCTG 1140
CAAACCTCTT ATGTCTTCCT GGGTATTGTT TTAGCCCTGA TAGGCGCTAT TTTCCTCCTG 1200
GTTTTGTATT TGAACCGCAA GGGGATAAAA AAGTGGATGC ATAACATCAG AGATGCCTGC 1260
AGGGATCACA TGGAAGGGTA TCATTACAGA TATGAAATCA ATGCGGACCC CAGATTAACA 1320
AACCTCAGTT CTAACTCGGA TGTCTGAGAA ATATTAGAGG ACAGACCAAG GACAACTCTG 1380
CATGAGATGT AGACTTAAGC TTTATCCCTA CTAGGCTTGC TCCACTTTCA TCCTCCACTA 1440
TAGATACAAC GGACTTTGAC TAAAAGCAGT GAAGGGGATT TGCTTCCTTG TTATGTAAAG 1500
TTTCTCGGTG TGTTCTGTTA ATGTAAGACG ATGAACAGTT GTGTATAGTG TTTTACCCTC 1560
TTCTTTTTCT TGGAACTCCT CAACACGTAT GGAGGGATTT TTCAGGTTTC AGCATGAACA 1620
TGGGCTTCTT GCTGTCTGTC TCTCTCTCAG TACAGTTCAA GGTGTAGCAA GTGTACCCAC 1680
ACAGATAGCA TTCAACAAAA GCTGCCTCAA CTTTTTCGAG AAAAATACTT TATTCATAAA 1740
TATCAGTTTT ATTCTCATGT ACCTAAGTTG TGGAGAAAAT AATTGCATCC TATAAACTGC 1800
CTGCAGACGT TAGCAGGCTC TTCAAAATAA CTCCATGGTG CACAGGAGCA CCTGCATCCA 1860
AGAGCATGCT TACATTTTAC TGTTCTGCAT ATTACAAAAA ATAACTTGCA ACTTCATAAC 1920
TTCTTTGACA AAGTAAATTA CTTTTTTGAT TGCAGTTTAT ATGAAAATGT ACTGATTTTT 1980
TTTTAATAAA CTGCATCGAG ATCCAACCGA CTGAATTGTT AAAAAAAAAA AAAAATAAAG 2040
ATTCTTAAAA GAA

BCA7 Protein sequence
Gene name: 5T4 oncofetal trophoblast glycoprotein; Unigene number: Hs.82128; Probeset Accession #: Z29083; Protein Accession #: NP_006661; Predicted Signal sequence: 1-32; Predicted TM domains: 357-373; PFAM domains: leucine_rich_repeats: 61-90, 119-142, 143-166, 235-258, 259-282, 294-345;
Summary: a type 1a TM protein of unknown function, detected in multiple cancers, with highest expression in breast cancer.

MPGGCSRGPA AGDGRLRLAR LALVLLGWVS SSSPTSSASS FSSSAPFLAS AVSAQPPLPD 60
QCPALCECSE AARTVKCVNR NLTEVPTDLP AYVRNLFLTG NQLAVLPAGA FARRPPLAEL 120
AALNLSGSRL DEVRAGAFEH LPSLRQLDLS HNPLADLSPF AFSGSNASVS APSPLVELIL 180
NHIVPPEDER QNRSFEGMVV AALLAGRALQ GLRRLELASN HFLYLPRDVL AQLPSLRHLD 240
LSNNSLVSLT YVSFRNLTHL ESLHLEDNAL KVLHNGTLAE LQGLPHIRVF LDNNPWVCDC 300
HMADMVTWLK ETEVVQGKDR LTCAYPEKMR NRVLLELNSA DLDCDPILPP SLQTSYVFLG 360
IVLALIGAIF LLVLYLNRKG IKKWMHNIRD ACRDHMEGYH YRYEINADPR LTNLSSNSDV

BCX5 DNA sequence
Gene name: LNIR; Unigene number: Hs.61460; Probeset Accession #: AA028028; Nucleic Acid Accession #: AF160477; Coding sequence: 225-1757 (start and stop codons underlined)

GGGGAGCTCG GAGCTCCCGA TCACGGCTTC TTGGGGGTAG CTACGGCTGG GTGTGTAGAA 60
CGGGGCCGGG GCTGGGGCTG GGTCCCCTAG TGAGACCCAA GTGCGAGAGG CAAGAACTCT 120
GCAGCTTCCT GCCTTCTGGG TCAGTTCCTT ATTCAAGTCT GCAGCCGGCT CCCAGGGAGA 180
TCTCGGTGGA ACTTCAGAAA CGCTGGGCAG TCTGCCTTTC AACCATGCCC CTGTCCCTGG 240
GAGCCGAGAT GTGGGGGCCT GAGGCCTGGC TGCTGCTGCT GCTACTGCTG GCATCATTTA 300
CAGGCCGGTG CCCCGCGGGT GAGCTGGAGA CCTCAGACGT GGTAACTGTG GTGCTGGGCC 360
AGGACGCAAA ACTGCCCTGC TTCTACCGAG GGGACTCCGG CGAGCAAGTG GGGCAAGTGG 420
CATGGGCTCG GGTGGACGCG GGCGAAGGCG CCCAGGAACT AGCGCTACTG CACTCCAAAT 480
ACGGGCTTCA TGTGAGCCCG GCTTACGAGG GCCGCGTGGA GCAGCCGCCG CCCCCACGCA 540
ACCCCCTGGA CGGCTCAGTG CTCCTGCGCA ACGCAGTGCA GGCGGATGAG GGCGAGTACG 600
AGTGCCGGGT CAGCACCTTC CCCGCCGGCA GCTTCCAGGC GCGGCTGCGG CTCCGAGTGA 660
TGGTGCCTCC CCTGCCCTCA CTGAATCCTG GTCCAGCACT AGAAGAGGGC CAGGGCCTGA 720
CCCTGGCAGC CTCCTGCACA GCTGAGGGCA GCCCAGCCCC CAGCGTGACC TGGGACACGG 780
AGGTCAAAGG CACAACGTCC AGCCGTTCCT TCAAGCACTC CCGCTCTGCT GCCGTCACCT 840
CAGAGTTCCA CTTGGTGCCT AGCCGCAGCA TGAATGGGCA GCCACTGACT TGTGTGGTGT 900
CCCATCCTGG CCTGCTCCAG GACCAAAGGA TCACCCACAT CCTCCACGTG TCCTTCCTTG 960
CTGAGGCCTC TGTGAGGGGC CTTGAAGACC AAAATCTGTG GCACATTGGC AGAGAAGGAG 1020
CTATGCTCAA GTGCCTGAGT GAAGGGCAGC CCCCTCCCTC ATACAACTGG ACACGGCTGG 1080
ATGGGCCTCT GCCCAGTGGG GTACGAGTGG ATGGGGACAC TTTGGGCTTT CCCCCACTGA 1140
CCACTGAGCA CAGCGGCATC TACGTCTGCC ATGTCAGCAA TGAGTTCTCC TCAAGGGATT 1200
CTCAGGTCAC TGTGGATGTT CTTGACCCCC AGGAAGACTC TGGGAAGCAG GTGGACCTAG 1260
TGTCAGCCTC GGTGGTGGTG GTGGGTGTGA TCGCCGCACT CTTGTTCTGC CTTCTGGTGG 1320
TGGTGGTGGT GCTCATGTCC CGATACCATC GGCGCAAGGC CCAGCAGATG ACCCAGAAAT 1380
ATGAGGAGGA GCTGACCCTG ACCAGGGAGA ACTCCATCCG GAGGCTGCAT TCCCATCACA 1440
CGGACCCCAG GAGCCAGCCG GAGGAGAGTG TAGGGCTGAG AGCCGAGGGC CACCCTGATA 1500
GTCTCAAGGA CAACAGTAGC TGCTCTGTGA TGAGTGAAGA GCCCGAGGGC CGCAGTTACT 1560
CCACGCTGAC CACGGTGAGG GAGATAGAAA CACAGACTGA ACTGCTGTCT CCAGGCTCTG 1620
GGCGGGCCGA GGAGGAGGAA GATCAGGATG AAGGCATCAA ACAGGCCATG AACCATTTTG 1680
TTCAGGAGAA TGGGACCCTA CGGGCCAAGC CCACGGGCAA TGGCATCTAC ATCAATGGGC 1740
GGGGACACCT GGTCTGACCC AGGCCTGCCT CCCTTCCCTA GGCCTGGCTC CTTCTGTTGA 1800
CATGGGAGAT TTTAGCTCAT CTTGGGGGCC TCCTTAAACA CCCCCATTTC TTGCGGAAGA 1860
TGCTCCCCAT CCCACTGACT GCTTGACCTT TACCTCCAAC CCTTCTGTTC ATCGGGAGGG 1920
CTCCACCAAT TGAGTCTCTC CCACCATGCA TGCAGGTCAC TGTGTGTGTG CATGTGTGCC 1980
TGTGTGAGTG TTGACTGACT GTGTGTGTGT GGAGGGGTGA CTGTCCGTGG AGGGGTGACT 2040
GTGTCCGTGG TGTGTATTAT GCTGTCATAT CAGAGTCAAG TGAACTGTGG TGTATGTGCC 2100
ACGGGATTTG AGTGGTTGCG TGGGCAACAC TGTCAGGGTT TGGCGTGTGT GTCATGTGGC 2160
TGTGTGTGAC CTCTGCCTGA AAAAGCAGGT ATTTTCTCAG ACCCCAGAGC AGTATTAATG 2220
ATGCAGAGGT TGGAGGAGAG AGGTGGAGAC TGTGGCTCAG ACCCAGGTGT GCGGGCATAG 2280
CTGGAGCTGG AATCTGCCTC CGGTGTGAGG GAACCTGTCT CCTACCACTT CGGAGCCATG 2340
GGGGCAAGTG TGAAGCAGCC AGTCCCTGGG TCAGCCAGAG GCTTGAACTG TTACAGAAGC 2400
CCTCTGCCCT CTGGTGGCCT CTGGGCCTGC TGCATGTACA TATTTTCTGT AAATATACAT 2460
GCGCCGGGAG CTTCTTGCAG GAATACTGCT CCGAATCACT TTTAATTTTT TTCTTTTTTT 2520
TTTCTTGCCC TTTCCATTAG TTGTATTTTT TATTTATTTT TATTTTTATT TTTTTTTAGA 2580
GATGGAGTCT CACTATGTTG CTCAGGCTGG CCTTGAACTC CTGGGCTCAA GCAATCCTCC 2640
TGCCTCAGCC TCCCTAGTAG CTGGGACTTT AAGTGTACAC CACTGTGCCT GCTTTGAATC 2700
CTTTACGAAG AGAAAAAAAA AATTAAAGAA AGCCTTTAGA TTTATCCAAT GTTTACTACT 2760
GGGATTGCTT AAAGTGAGGC CCCTCCAACA CCAGGGGGTT AATTCCTGTG ATTGTGAAAG 2820
GGGCTACTTC CAAGGCATCT TCATGCAGGC AGCCCCTTGG GAGGGCACCT GAGAGCTGGT 2880
AGAGTCTGAA ATTAGGGATG TGAGCCTCGT GGTTACTGAG TAAGGTAAAA TTGCATCCAC 2940
CATTGTTTGT GATACCTTAG GGAATTGCTT GGACCTGGTG ACAAGGGCTC CTGTTCAATA 3000
GTGGTGTTGG GGAGAGAGAG AGCAGTGATT ATAGACCGAG AGAGTAGGAG TTGAGGTGAG 3060
GTGAAGGAGG TGCTGGGGGT GAGAATGTCG CCTTTCCCCC TGGGTTTTGG ATCACTAATT 3120
CAAGGCTCTT CTGGATGTTT CTCTGGGTTG GGGCTGGAGT TCAATGAGGT TTATTTTTAG 3180
CTGGCCCACC CAGATACACT CAGCCAGAAT ACCTAGATTT AGTACCCAAA CTCTTCTTAG 3240
TCTGAAATCT GCTGGATTTC TGGCCTAAGG GAGAGGCTCC CATCCTTCGT TCCCCAGCCA 3300
GCCTAGGACT TCGAATGTGG AGCCTGAAGA TCTAAGATCC TAACATGTAC ATTTTATGTA 3360
AATATGTGCA TATTTGTACA TAAAATGATA TTCTGTTTTT AAATAAACAG ACAAAACTTG 3420
TTCAAAAAAA AAAAAAAAAA AAAAAAAAA

BCX5 Protein sequence
Gene name: LNIR; Unigene number: Hs.61460; Probeset Accession #: AA028028; Protein Accession #: AF160477; Predicted Signal sequence: 1-26; Predicted TM domains: 355-371; PFAM domains: IgSF_domain: 45-129, 162-225, 263-317; Summary: A type Ia TM protein; is a member of the immunoglobulin superfamily.

MPLSLGAEMW GPEAWLLLLL LLASFTGRCP AGELETSDVV TVVLGQDAKL PCFYRGDSGE 60
QVGQVAWARV DAGEGAQELA LLHSKYGLHV SPAYEGRVEQ PPPPRNPLDG SVLLRNAVQA 120
DEGEYECRVS TFPAGSFQAR LRLRVMVPPL PSLNPGPALE EGQGLTLAAS CTAEGSPAPS 180
VTWDTEVKGT TSSRSFKHSR SAAVTSEFHL VPSRSMNGQP LTCVVSHPGL LQDQRITHIL 240
HVSFLAEASV RGLEDQNLWH IGREGAMLKC LSEGQPPPSY NWTRLDGPLP SGVRVDGDTL 300
GFPPLTTEHS GIYVCHVSNE FSSRDSQVTV DVLDPQEDSG KQVDLVSASV VVVGVIAALL 360

FCLLVVVVVL MSRYHRRKAQ QMTQKYEEEL TLTRENSIRR LHSHHTDPRS QPEESVGLRA 420
EGHPDSLKDN SSCSVMSEEP EGRSYSTLTT VREIETQTEL LSPGSGRAEE EEDQDEGIKQ 480
AMNHFVQENG TLRAKPTGNG IYINGRGHLV

mouse BCX5 Protein sequence
Gene name: mouse_LNIR; Unigene number: n/a; Probeset Accession #: BF168327; Protein Accession #: n/a; Predicted Signal sequence: 1-27; Predicted TM domains: 346-362; PFAM domains: IgSF_domains:44-126,166-221,259-313; Summary: This is the mouse orthologue of human BCX5; it is a type 1a TM protein of unknown function.

MPLSLGAEMW GPEAWLRLLF LASFTGQYSA GELETSDVVT VVLGQDAKLP CFYRGDPDEQ 60
VGQVAWARVD PNEXYPGAGL LHSKYGLHVN PAYEDRVEQX XHETFRRSVL LRNAVQADEG 120
EYECRVSTFP SGSFQARMRL RVLVPPLPSL NPGPPLEEGQ ADVAASCTAE GSPAPSVTWD 180
TEVKGTQSSR SFTHPRSAAV TSEFHLVPSR SMNGQPLTCV VSHPGLLQDR RITHTLQVAF 240
LAEASVRGLE DQNLWQVGRE GATLKCLSEG QPPPKYNWTR LDGPLPSGVR VKGDTLGFPP 300
LTTEHSGVYX CHVSNELSSR DSQVTVEVLD PEDPGKQVDL VSASVIIVGV IAALLFCLLV 360
VVVVLMSRYH RRKAQQMTQK YEEELTLTRE NSIRRLHSHH SDPRSQPEES VGLRAEGHPD 420
SLKDNSSCSV MSEEPEGRSY STLTTVREIE TQTELLSPGS GRTEEDDDQD EGIKQAMNHL 480
CRKMGP

BCZ6 DNA sequence
Gene name: IL-6 receptor beta chain (gp130; oncostatin M receptor); Unigene number: Hs.82065; Probeset Accession #: M57230 / AA406546; Nucleic Acid Accession #: NM_002184; Coding sequence: 256-3012 (start and stop codons underlined)

GAGCAGCCAA AAGGCCCGCG GAGTCGCGCT GGGCCGCCCC GGCGCAGCTG AACCGGGGGC 60
CGCGCCTGCC AGGCCGACGG GTCTGGCCCA GCCTGGCGCC AAGGGGTTCG TGCGCTGTGG 120
AGACGCGGAG GGTCGAGGCG GCGCGGCCTG AGTGAAACCC AATGGAAAAA GCATGACATT 180
TAGAAGTAGA AGACTTAGCT TCAAATCCCT ACTCCTTCAC TTACTAATTT TGTGATTTGG 240
AAATATCCGC GCAAGATGTT GACGTTGCAG ACTTGGGTAG TGCAAGCCTT GTTTATTTTC 300
CTCACCACTG AATCTACAGG TGAACTTCTA GATCCATGTG GTTATATCAG TCCTGAATCT 360
CCAGTTGTAC AACTTCATTC TAATTTCACT GCAGTTTGTG TGCTAAAGGA AAAATGTATG 420
GATTATTTTC ATGTAAATGC TAATTACATT GTCTGGAAAA CAAACCATTT TACTATTCCT 480
AAGGAGCAAT ATACTATCAT AAACAGAACA GCATCCAGTG TCACCTTTAC AGATATAGCT 540
TCATTAAATA TTCAGCTCAC TTGCAACATT CTTACATTCG GACAGCTTGA ACAGAATGTT 600
TATGGAATCA CAATAATTTC AGGCTTGCCT CCAGAAAAAC CTAAAAATTT GAGTTGCATT 660
GTGAACGAGG GGAAGAAAAT GAGGTGTGAG TGGGATGGTG GAAGGGAAAC ACACTTGGAG 720
ACAAACTTCA CTTTAAAATC TGAATGGGCA ACACACAAGT TTGCTGATTG CAAAGCAAAA 780
CGTGACACCC CCACCTCATG CACTGTTGAT TATTCTACTG TGTATTTTGT CAACATTGAA 840
GTCTGGGTAG AAGCAGAGAA TGCCCTTGGG AAGGTTACAT CAGATCATAT CAATTTTGAT 900
CCTGTATATA AAGTGAAGCC CAATCCGCCA CATAATTTAT CAGTGATCAA CTCAGAGGAA 960
CTGTCTAGTA TCTTAAAATT GACATGGACC AACCCAAGTA TTAAGAGTGT TATAATACTA 1020
AAATATAACA TTCAATATAG GACCAAAGAT GCCTCAACTT GGAGCCAGAT TCCTCCTGAA 1080
GACACAGCAT CCACCCGATC TTCATTCACT GTCCAAGACC TTAAACCTTT TACAGAATAT 1140
GTGTTTAGGA TTCGCTGTAT GAAGGAAGAT GGTAAGGGAT ACTGGAGTGA CTGGAGTGAA 1200
GAAGCAAGTG GGATCACCTA TGAAGATAGA CCATCTAAAG CACCAAGTTT CTGGTATAAA 1260
ATAGATCCAT CCCATACTCA AGGCTACAGA ACTGTACAAC TCGTGTGGAA GACATTGCCT 1320
CCTTTTGAAG CCAATGGAAA AATCTTGGAT TATGAAGTGA CTCTCACAAG ATGGAAATCA 1380
CATTTACAAA ATTACACAGT TAATGCCACA AAACTGACAG TAAATCTCAC AAATGATCGC 1440
TATCTAGCAA CCCTAACAGT AAGAAATCTT GTTGGCAAAT CAGATGCAGC TGTTTTAACT 1500
ATCCCTGCCT GTGACTTTCA AGCTACTCAC CCTGTAATGG ATCTTAAAGC ATTCCCCAAA 1560
GATAACATGC TTTGGGTGGA ATGGACTACT CCAAGGGAAT CTGTAAAGAA ATATATACTT 1620
GAGTGGTGTG TGTTATCAGA TAAAGCACCC TGTATCACAG ACTGGCAACA AGAAGATGGT 1680
ACCGTGCATC GCACCTATTT AAGAGGGAAC TTAGCAGAGA GCAAATGCTA TTTGATAACA 1740
GTTACTCCAG TATATGCTGA TGGACCAGGA AGCCCTGAAT CCATAAAGGC ATACCTTAAA 1800
CAAGCTCCAC CTTCCAAAGG ACCTACTGTT CGGACAAAAA AAGTAGGGAA AAACGAAGCT 1860
GTCTTAGAGT GGGACCAACT TCCTGTTGAT GTTCAGAATG GATTTATCAG AAATTATACT 1920
ATATTTTATA GAACCATCAT TGGAAATGAA ACTGCTGTGA ATGTGGATTC TTCCCACACA 1980
GAATATACAT TGTCCTCTTT GACTAGTGAC ACATTGTACA TGGTACGAAT GGCAGCATAC 2040
ACAGATGAAG GTGGGAAGGA TGGTCCAGAA TTCACTTTTA CTACCCCAAA GTTTGCTCAA 2100
GGAGAAATTG AAGCCATAGT CGTGCCTGTT TGCTTAGCAT TCCTATTGAC AACTCTTCTG 2160
GGAGTGCTGT TCTGCTTTAA TAAGCGAGAC CTAATTAAAA AACACATCTG GCCTAATGTT 2220
CCAGATCCTT CAAAGAGTCA TATTGCCCAG TGGTCACCTC ACACTCCTCC AAGGCACAAT 2280
TTTAATTCAA AAGATCAAAT GTATTCAGAT GGCAATTTCA CTGATGTAAG TGTTGTGGAA 2340
ATAGAAGCAA ATGACAAAAA GCCTTTTCCA GAAGATCTGA AATCATTGGA CCTGTTCAAA 2400
AAGGAAAAAA TTAATACTGA AGGACACAGC AGTGGTATTG GGGGGTCTTC ATGCATGTCA 2460
TCTTCTAGGC CAAGCATTTC TAGCAGTGAT GAAAATGAAT CTTCACAAAA CACTTCGAGC 2520
ACTGTCCAGT ATTCTACCGT GGTACACAGT GGCTACAGAC ACCAAGTTCC GTCAGTCCAA 2580
GTCTTCTCAA GATCCGAGTC TACCCAGCCC TTGTTAGATT CAGAGGAGCG GCCAGAAGAT 2640
CTACAATTAG TAGATCATGT AGATGGCGGT GATGGTATTT TGCCCAGGCA ACAGTACTTC 2700
AAACAGAACT GCAGTCAGCA TGAATCCAGT CCAGATATTT CACATTTTGA AAGGTCAAAG 2760
CAAGTTTCAT CAGTCAATGA GGAAGATTTT GTTAGACTTA AACAGCAGAT TTCAGATCAT 2820

ATTTCACAAT CCTGTGGATC TGGGCAAATG AAAATGTTTC AGGAAGTTTC TGCAGCAGAT 2880
GCTTTTGGTC CAGGTACTGA GGGACAAGTA GAAAGATTTG AAACAGTTGG CATGGAGGCT 2940
GCGACTGATG AAGGCATGCC TAAAAGTTAC TTACCACAGA CTGTACGGCA AGGCGGCTAC 3000
ATGCCTCAGT GAAGGACTAG TAGTTCCTGC TACAACTTCA GCAGTACCTA TAAAGTAAAG 3060
CTAAAATGAT TTTATCTGTG AATTC

BCZ6 Protein sequence
Gene name: IL-6 receptor beta chain (gp130; oncostatin M receptor); Unigene number: Hs.82065; Probeset Accession #: M57230 / AA406546; Protein Accession #: NP_002175; Predicted Signal sequence: 1-22; Predicted TM domains: 625-641; PFAM domains: fibronectin_type_III_domains: 222-311, 424-509, 519-606; Summary: A type I TM protein; it homodimerizes or heterodimerized to make a functional receptor for IL-6, oncostatin-M, IL-11, LIF, and CNTF.

MLTLQTWVVQ ALFIFLTTES TGELLDPCGY ISPESPVVQL HSNFTAVCVL KEKCMDYFHV 60
NANYIVWKTN HFTIPKEQYT IINRTASSVT FTDIASLNIQ LTCNILTFGQ LEQNVYGITI 120
ISGLPPEKPK NLSCIVNEGK KMRCEWDGGR ETHLETNFTL KSEWATHKFA DCKAKRDTPT 180
SCTVDYSTVY FVNIEVWVEA ENALGKVTSD HINFDPVYKV KPNPPHNLSV INSEELSSIL 240
KLTWTNPSIK SVIILKYNIQ YRTKDASTWS QIPPEDTAST RSSFTVQDLK PFTEYVFRIR 300
CMKEDGKGYW SDWSEEASGI TYEDRPSKAP SFWYKIDPSH TQGYRTVQLV WKTLPPFEAN 360
GKILDYEVTL TRWKSHLQNY TVNATKLTVN LTNDRYLATL TVRNLVGKSD AAVLTIPACD 420
FQATHPVMDL KAFPKDNMLW VEWTTPRESV KKYILEWCVL SDKAPCITDW QQEDGTVHRT 480
YLRGNLAESK CYLITVTPVY ADGPGSPESI KAYLKQAPPS KGPTVRTKKV GKNEAVLEWD 540
QLPVDVQNGF IRNYTIFYRT IIGNETAVNV DSSHTEYTLS SLTSDTLYMV RMAAYTDEGG 600
KDGPEFTFTT PKFAQGEIEA IVVPVCLAFL LTTLLGVLFC FNKRDLIKKH IWPNVPDPSK 660
SHIAQWSPHT PPRHNFNSKD QMYSDGNFTD VSVVEIEAND KKPFPEDLKS LDLFKKEKIN 720
TEGHSSGIGG SSCMSSSRPS ISSSDENESS QNTSSTVQYS TVVHSGYRHQ VPSVQVFSRS 780
ESTQPLLDSE ERPEDLQLVD HVDGGDGILP RQQYFKQNCS QHESSPDISH FERSKQVSSV 840
NEEDFVRLKQ QISDHISQSC GSGQMKMFQE VSAADAFGPG TEGQVERFET VGMEAATDEG 900
MPKSYLPQTV RQGGYMPQ

BFG4 DNA sequence
Gene name: KIAA0882 protein; Unigene number: Hs.90419; Probeset Accession #: Z39762; Nucleic Acid Accession #: AB020689; Coding sequence: 108-2777 (start and stop codons underlined)

GAACTTATGT AGCCTCATTA TCCCGCTCCG TGAGGTGACA ATTGTGGAAA AGGCAGACAG 60
CTCCAGTGTG CTCCCCAGTC CCTTATCACA TCAGCACCCG AAACAGGATG ACCTTCCTAT 120
TTGCCAACTT GAAAGATAGA GACTTTCTAG TGCAGAGGAT CTCAGATTTC CTGCAACAGA 180
CTACTTCCAA AATATATTCT GACAAGGAGT TTGCAGGAAG TTACAACAGT TCAGATGATG 240
AGGTGTACTC TCGACCCAGC AGCCTCGTCT CCTCCAGCCC CCAGAGAAGC ACGAGCTCTG 300
ATGCTGATGG AGAGCGCCAG TTTAACCTAA ATGGCAACAG CGTCCCCACA GCCACACAGA 360
CCCTGATGAC CATGTATCGG CGGCGGTCTC CCGAGGAGTT CAACCCGAAA TTGGCCAAAG 420
AGTTTCTGAA AGAGCAAGCC TGGAAGATTC ACTTTGCTGA GTATGGGCAA GGGATCTGCA 480
TGTACCGCAC AGAGAAAACG CGGGAGCTGG TGTTGAAGGG CATCCCGGAG AGCATGCGTG 540
GGGAGCTCTG GCTGCTGCTG TCAGGTGCCA TCAATGAGAA GGCCACACAT CCTGGGTACT 600
ATGAAGACCT AGTGGAGAAG TCCATGGGGA AGTATAATCT CGCCACGGAG GAGATTGAGA 660
GGGATTTACA CCGCTCCCTT CCAGAACACC CAGCTTTTCA GAATGAAATG GGCATTGCTG 720
CACTAAGGAG AGTCTTAACA GCTTATGCTT TTCGAAATCC CAACATAGGG TATTGCCAGG 780
CCATGAATAT TGTCACTTCA GTGCTGCTGC TTTATGCCAA AGAGGAGGAA GCTTTCTGGC 840
TGCTTGTGGC TTTGTGTGAG CGCATGCTCC CAGATTACTA CAACACCAGA GTTGTGGGTG 900
CACTGGTGGA CCAAGGTGTC TTTGAGGAGC TAGCACGAGA CTACGTCCCA CAGCTGTACG 960
ACTGCATGCA AGACCTGGGC GTGATTTCCA CCATCTCCCT GTCTTGGTTC CTCACACTAT 1020
TTCTCAGTGT GATGCCTTTT GAGAGTGCAG TTGTGGTTGT TGACTGTTTC TTCTATGAAG 1080
GAATTAAAGT GATATTCCAG TTGGCCCTAG CTGTGCTGGA TGCAAATGTG GACAAACTGT 1140
TGAACTGCAA GGATGATGGG GAGGCCATGA CCGTTTTGGG AAGGTATTTA GACAGTGTGA 1200
CCAATAAAGA CAGCACACTG CCTCCCATTC CTCACCTCCA CTCCTTGCTC AGCGATGATG 1260
TGGAACCTTA CCCTGAGGTA GACATCTTTA GACTCATCAG AACTTCCTAC GAGAAATTCG 1320
GAACTATCCG GGCAGATTTG ATTGAACAGA TGAGATTCAA ACAGAGACTG AAAGTGATCC 1380
AGACGCTGGA GGATACTACG AAACGCAACG TGGTACGAAC CATTGTGACA GAAACTTCCT 1440
TTACCATTGA TGAGCTGGAA GAACTTTATG CTCTTTTCAA GGCAGAACAT CTCACCAGCT 1500
GCTACTGGGG CGGGAGCAGC AACGCGCTGG ACCGGCATGA CCCCAGCCTG CCCTACCTGG 1560
AACAGTATCG CATTGACTTC GAGCAGTTCA AGGGAATGTT TGCTCTTCTC TTTCCTTGGG 1620
CATGTGGAAC TCACTCTGAC GTTCTGGCCT CCCGCTTGTT CCAGTTATTA GATGAAAATG 1680
GAGACTCTTT GATTAACTTC CGGGAGTTTG TCTCTGGGCT AAGTGCTGCA TGCCATGGGG 1740
ACCTCACAGA GAAGCTCAAA CTCCTGTACA AAATGCACGT CTTGCCTGAG CCATCCTCTG 1800
ATCAAGATGA ACCAGATTCT GCTTTTGAAG CAACTCAGTA CTTCTTTGAA GATATTACCC 1860
CAGAATGTAC ACATGTTGTT GGATTGGATA GCAGAAGCAA ACAGGGTGCA GATGATGGCT 1920
TTGTTACGGT GAGCCTAAAG CCAGACAAAG GGAAGAGAGC AAATTCCCAA GAAAATCGTA 1980
ATTATTTGAG ACTGTGGACT CCAGAAAATA AATCTAAGTC AAAGAATGCA AAGGATTTAC 2040
CCAAATTAAA TCAGGGGCAG TTCATTGAAC TGTGTAAGAC AATGTATAAC ATGTTCAGCG 2100
AAGACCCCAA TGAGCAGGAG CTGTACCATG CCACGGCAGC AGTGACCAGC CTCCTGCTGG 2160

AGATTGGGGA GGTCGGCAAG TTGTTCGTGG CCCAGCCTGC AAAGGAGGGC GGGAGCGGAG 2220
GCAGTGGGCC GTCCTGCCAC CAGGGCATCC CAGGCGTGCT CTTCCCCAAG AAAGGGCCAG 2280
GCCAGCCTTA CGTGGTGGAG TCTGTTGAGC CCCTGCCGGC CAGCCTGGCC CCCGACAGCG 2340
AGGAACACTC CCTTGGAGGA CAAATGGAGG ACATCAAGCT GGAGGACTCC TCGCCCCGGG 2400
ACAACGGGGC CTGCTCCTCC ATGCTGATCT CTGACGACGA CACCAAGGAC GACAGCTCCA 2460
TGTCCTCATA CTCGGTGCTG AGTGCCGGCT CCCACGAGGA GGACAAGCTG CACTGCGAGG 2520
AAATCGGAGA GGACACGGTC CTGGTGCGGA GCGGCCAGGG CACGGCGGCA CTGCCCCGGA 2580
GCACCAGCCT GGACCGGGAC TGGGCCATCA CCTTCGAGCA GTTCCTGGCC TCCCTCTTAA 2640
CTGAGCCTGC CCTGGTCAAG TACTTTGACA AGCCCGTGTG CATGATGGCC AGGATTACCA 2700
GTGCAAAAAA CATCCGGATG ATGGGCAAGC CCCTCACCTC GGCCAGTGAC TATGAAATCT 2760
CGGCCATGTC CGGCTGACAC GGGCGCCTTC CCGGGGGAGT GGGAGGAGAG GGAGGGGAGG 2820
GATTTTTTAT GTTCTTCTGT GTTGAGTTTT TTCTTTCTTT CTTTTAAATT AAATATTTAT 2880
TAGTACCTGG AATTGAAGCC TAGTGTTTTC ATAATGTAAT TCAATGAAAA CTGTTGGAGA 2940
AATATTTAAA CACCTCAATG TAGGTACATT ACACTCTTGT TGCGGGGAGG GGATTTACCA 3000
GAATACAGTT TATTTCGTGA ATTCTAAAAA ACAAAAAGAT GAATCTGTCA GTGATATGTG 3060
TGTATTATAA CTTATTAATC TTGCTGTTGA GCTGTATACA TGGTTTAAAA AATAGTACTG 3120
TTTAATGCTA AGTAAGGCAG CAGTCATTTG TGTATTCAGG CTTTTTAAAT AAAATTAGAG 3180
CTGTAAGGAA AATGAAAAGC CACAAATGCA AGACTGTTCT TAAATGGAAG GCATAGTCAG 3240
CGAGGGTAAA TCCTATACCA CTTTAGGAAG TATTAAAAAT ATTTTTAAGA TTTGAAATAT 3300
ATTTCATAGA AGTCCTCTAT TCAAAATCAT ATTCCACAGA TGTTCCCCTT CAAAGGGAAA 3360
ACATTTGGGG TTCTAAACAG TTATGAAAGT AAGTGATTTT TACATGATTC CAGAATAACA 3420
CTTGTATTGA CCAATTTAGA CAGATACCAG ACCAATTTTG CATTTAAGAA ATTGTTCTGA 3480
TTATTTACGT CAACTCATTA GAATTCAGTG AAAAGTAACA GTCTTTTGTC ACAGAGAATC 3540
TGAAAGTAGC AGCAAAGACA GAGGGCTCAT GACAGGTTTT TGCTTTTGCT TTGCTTTTGT 3600
TTTTGAAAGA GTAAAAGTAC TGATGCTTCT GATACTGGAT GTTTAGCTTC TTACTGCAAA 3660
AACATAAGTA AAACAGTCAA CTTTACCATT TCCGTATTCT CCATAGATTG AAGAAATTTA 3720
TACCACATAT CGCATATGAC CATCTTTCCA TCAAATCAAT GTAGAGATAA TGTAAACTGA 3780
AAAAAAATCT GCAAGATAAT GTAACTGAAT GTTTTAAAAA CAGAACTTGT CACTTTATAT 3840
AAAAGAATAG TATGCTCTAT TTCCTGAATG GATGTGGAAA TGAAAGCTAG CGCACCTGCA 3900
CTTTGAATTC TTGCTTCTTT TTTATTACTG TTATGATTTT GCTTTTTACA GATGTTGGAC 3960
GATTTTTTCT TCTGATTGTT GAATTCATAA TCATGGTCTC ATTTCCTTTG CTTCTTTGGA 4020
ATATTTCTTT CAACACATTC CTTTATTTTA TTATACATTG TGTCCTTTTT TTAGCTATTG 4080
CTGCTGTTGT TTTTTATTCT ATTTACAGGA TGATTTTTAA ACTGTCAAAT GAAGTAGTGT 4140
TAACCTCAAA TAGGCTAAAT GTGAACAAAT AAAATACAGC AAATACTCAG AAAAAAAAAA 4200
AAAAAAAAAA AAAAA

BFG4 Protein sequence
Gene name: KIAA0882 protein; Unigene number: Hs.90419; Probeset Accession #: Z39762; Protein Accession #: BAA74905; Signal sequence: none; Predicted TM domains: 302-318; PFAM domains: TBC_domain: 135-347; Summary: a Type II membrane protein, likely localized to the peroxisome.

MTFLFANLKD RDFLVQRISD FLQQTTSKIY SDKEFAGSYN SSDDEVYSRP SSLVSSSPQR 60
STSSDADGER QFNLNGNSVP TATQTLMTMY RRRSPEEFNP KLAKEFLKEQ AWKIHFAEYG 120
QGICMYRTEK TRELVLKGIP ESMRGELWLL LSGAINEKAT HPGYYEDLVE KSMGKYNLAT 180
EEIERDLHRS LPEHPAFQNE MGIAALRRVL TAYAFRNPNI GYCQAMNIVT SVLLLYAKEE 240
EAFWLLVALC ERMLPDYYNT RVVGALVDQG VFEELARDYV PQLYDCMQDL GVISTISLSW 300
FLTLFLSVMP FESAVVVVDC FFYEGIKVIF QLALAVLDAN VDKLLNCKDD GEAMTVLGRY 360
LDSVTNKDST LPPIPHLHSL LSDDVEPYPE VDIFRLIRTS YEKFGTIRAD LIEQMRFKQR 420
LKVIQTLEDT TKRNVVRTIV TETSFTIDEL EELYALFKAE HLTSCYWGGS SNALDRHDPS 480
LPYLEQYRID FEQFKGMFAL LFPWACGTHS DVLASRLFQL LDENGDSLIN FREFVSGLSA 540
ACHGDLTEKL KLLYKMHVLP EPSSDQDEPD SAFEATQYFF EDITPECTHV VGLDSRSKQG 600
ADDGFVTVSL KPDKGKRANS QENRNYLRLW TPENKSKSKN AKDLPKLNQG QFIELCKTMY 660
NMFSEDPNEQ ELYHATAAVT SLLLEIGEVG KLFVAQPAKE GGSGGSGPSC HQGIPGVLFP 720
KKGPGQPYVV ESVEPLPASL APDSEEHSLG GQMEDIKLED SSPRDNGACS SMLISDDDTK 780
DDSSMSSYSV LSAGSHEEDK LHCEEIGEDT VLVRSGQGTA ALPRSTSLDR DWAITFEQFL 840
ASLLTEPALV KYFDKPVCMM ARITSAKNIR MMGKPLTSAS DYEISAMSG

BCU7 DNA sequence
Gene name: EST; Unigene number: Hs.98558; Probeset Accession #: AA428062; Nucleic Acid Accession #: n/a; Coding sequence: 1-573 (stop codon underlined)

TATTTTATTT TCCAGGCTAA AGCAAATGAA AGTTTGCTGG TATCAACACA GCCTGCCATA 60
TTTTTCACAG CATGCAACAA TGGTGCTAGG ATAGCTATTT CTTACTGTAA TTGCCAGAGG 120
CAGAAATGGT CTGGGTATAA GCTATTTCAT AAAAGCAGCT TTAAATTGTC AGTATTAAGG 180
TTTTCATGTG GAAAGGTGTC ATTCAAAAAA AAAGTAATTG GCATACATAT TCCACATCAT 240
CGATCCTCTC TGTGGTGTTA ATTTTTTTAT ATGACCAGTA GAAAAATTTT AATATTCTCA 300
CAATATAGGT TTTGGGGCTT CCATATCATC AAAAGACTGA AAAATTATAA TTTTAGAATT 360
AAACTGATGG ATTTCATTAT AGAATTATCT GTGAGTTGTG TAGACACAGT CTTAATGTTT 420
CTGGTTATGA CAGATAAGTT TGCTCAAAAA ATGTGGATGA AGCCATTATT GTTATTATTG 480
TTATTGCTTC TGTTCAGTTG TCTAAGTATC ATCCCTTCTG TGGCCCATCA CGCAGCAGAG 540
TTGCCCTACA AATTTCATTT GGCAGCGCCA TAACATTCAT TTAAAAAGTT TATGAAAACA 600

TTCATTTGAA AGTTCCATGC AGCTTTAGCA CAGAGTTGAC CAAACACTGG CGTAAGTTCA 660
ATTTACACAG AATATTTGAA TTGAAACAAT AGAAATTTTT CTCATAATAT ATACCTATGT 720
GAAACCAACT TATCTGCATA ATTAAATCTA ATACATATTT AAGCCAGTTT AAGTGCTTTG 780
TGTTGATGCC ATGCTTATCA AATACATGCA CAAGCTAAAC ATAATTTGAA TGGGTCTATG 840
AAGGAAAAAT AATGCTTAGA CTTTGGTGTA GGTTCTTCCT GTGTAGCCAT ATACCCAGGC 900
TCTGCAGTAT CGAAGGATGC AAATGTTGAC ATAGATGGAA GCTCTTACCT ACCAAAGTGT 960
TTAGGAAGGA TAAAGTTACA TTTGTCTTAA TTTCTAACAT TATCTTTGCT TTTATGTTTC 1020
ATAAAAATTT GTCATTATTT ATGCTGGTGA AACGTATAAT CACATCCAAT TATTTGAACA 1080
CATGCAAAAT AATTTTTTAA ATTATGTTAT TGTTTAAATT TGACTTATGG GAGATCAGTC 1140
AAAAACTTAG AAGGTTTAAC ACTTCACTGA TTAATGGTGC TGAAAACACG TTACAATTAC 1200
CACATATCCT TGCTATAAGT TTTGAAGTTT CTTAGCAATT AAAGTTTTTT TATTCAGTGT 1260
GAACTGTCAG TATCTATTCT GGTGCTAAAT GTATGGTGCT AAATGAATTG TTAGTGTTGA 1320
TGGCTTTAGT AATGCTCCTT TTATTCATTG CTAAATTTAG TGTTATCCAT TTGATTCCTG 1380
ATTCAGAAAT ATCAATAAAA TCCTATGTTA AATTAATCTT TACCAAAAAC AGGCAAGTTA 1440
ACTCTGTTGT TTTAATTCAA CAGTCCAACA TTATTTAGGT GTTACAGAGT GTAAATATAT 1500
TTCTTTGGGA GTTATTTTCT TTTTAAAATC TTTTTATAGC TTGGCAATGT CCAAAGTCAA 1560
ATATCACCTA AACTGGTTAG ATTACTTCTA CAGCTAATAA TATTGCAGGC ACTGGCGCCC 1620
TCTGGTGGTT ATGAAGACAA ATTCTTAATG GCTACTTGAC CTACAGCAAA AGCCATTTCT 1680
GTACCATAAA AATTTGTTGT GCAATATTAG AATTATCATA TGTTTCCTAC ATCTGACAGC 1740
ACCTAAAATG TTTGATAATA TTAACATGTA TCTAAGAGGA AAAAAGAGTT AATATATTCT 1800
GGCACCCACT TTCCTAGTAA TGTTTTCCAT GATTTTCCAG TTCTGAGGCA CTTATTAAAG 1860
TGCTTTTTTT TTTCTGAATT AATTAGGTAT TGGTAAAATA TATTTTTAAA TTTAGTTAGC 1920
TTTATAAACA CAATTAGAAT TACAATTAAT TAACAGAGGT ATAATTGTCT CACTTTCAGA 1980
AGTGATCATT TATTTTTATT TAGCACAGGT CATAAGAAAA ATATATAGAA AAATAATCAA 2040
TTTCATATAT AAAAGGATTA TTTCTCCACC TTTAATTATT GGCCTATCAT TTGTTAGTGT 2100
TATTTGGTCA TATTATTGAA CTAATGTATT ATTCCATTCA AAGTCTTTCT AGATTTAAAA 2160
ATGTATGCAA AAGCTTAGGA TTATATCATG TGTAACTATT ATAGATAACA TCCTAAACCT 2220
TCAGTTTAGA TATATAATTG ACTGGGTGTA ATCTCTTTTG TAATCTGTTT TGACAGATTT 2280
CTTAAATTAT GTTAGCATAA TCAAGGAAGA TTTACCTTGA AGCACTTTCC AAATTGATAC 2340
TTTCAAACTT ATTTTAAAGC AGTAGAACCT TTTCTATGAA CTAAATCACA TGCAAAACTC 2400
CAACCTGTAG TATACATAAA ATGGACTTAC TTATTCCTCT CACCTTCTCC AGTGCCTAGG 2460
AATATTCTTC TCTGAGCCCT AGGATTGATT CTATCACACA GAGCAACATT AATCTAAATG 2520
GTTTAGCTCC CTCTTTTTTC TCTAAAAACA ATCAGCTAAT AAAAAAAAAA TTTGAGGGCC 2580
TAAATTATTT CAATGGTTGT TTGAAATATT CAGTTCAGTT TGTACCTGTT AGCAGTCTTT 2640
CAGTTTGGGG GAGAATTAAA TACTGTGCTA AGCTGGTGCT TGGATACATA TTACAGCATC 2700
TTGTGTTTTA TTTGACAAAC AGAATTTTGG TGCCATAATA TTTTGAGAAT TAGAGAAGAT 2760
TGTGATGCAT ATATATAAAC ACTATTTTTA AAAAATATCT AAATATGTCT CACATATTTA 2820
TATAATCCTC AAATATACTG TACCATTTTA GATATTTTTT AAACAGATTA ATTTGGAGAA 2880
GTTTTATTCA TTACCTAATT CTGTGGCAAA AATGGTGCCT CTGATGTTGT GATATAGTAT 2940
TGTCAGTGTG TACATATATA AAACCTGTGT AAACCTCTGT CCTTATGAAC CATAACAAAT 3060
GTAGCTTTTT AAAGTCCATT GTATTGTTTT TTCTTTCAAT AAAAGAGTAT AATTAATTGG 3060
TTGTTTTTGA

BCU7 Protein sequence
Gene name: EST; Unigene number: Hs.98558; Probeset Accession #: AA428062; Protein Accession #: n/a; Signal sequence: none; Predicted TM domains: 125-141, 154-170; PFAM domains: none; Summary: A type III membrane protein, highly overexpressed in breast cancer and prostate cancer; unknown function.

YFIFQAKANE SLLVSTQPAI FFTACNNGAR IAISYCNCQR QKWSGYKLFH KSSFKLSVLR 60
FSCGKVSFKK KVIGIHIPHH RSSLWCXFFY MTSRKILIFS QYRFWGFHII KRLKNYNFRI 120
KLMDFIIELS VSCVDTVLMF LVMTDKFAQK MWMKPLLLLL LLLLFSCLSI IPSVAHHAAE 180
LPYKFHLAAP

BFA1 DNA sequence
Gene name: calsyntenin-2; Unigene number: Hs.7413; Probeset Accession #: R46025; Nucleic Acid Accession #: NM_022131; Coding sequence: 11-2878 (start and stop codons underlined)

TGCTGCGAGG ATGCTGCCTG GGCGGCTGTG CTGGGTGCCG CTCCTGCTGG CGCTGGGCGT 60
GGGGAGCGGC AGCGGCGGTG GCGGGGACAG CCGGCAGCGC CGCCTCCTCG CGGCTAAAGT 120
CAATAAGCAC AAGCCATGGA TCGAGACTTC ATATCATGGA GTCATAACTG AGAACAATGA 180
CACAGTCATT TTGGACCCAC CACTGGTAGC CCTGGATAAA GATGCACCGG TTCCTTTTGC 240
AGGGGAAATC TGTGCGTTCA AGATCCATGG CCAGGAGCTG CCCTTTGAGG CTGTGGTGCT 300
CAACAAGACA TCAGGAGAGG GCCGGCTCCG TGCCAAGAGC CCCATTGACT GTGAGTTGCA 360
GAAGGAGTAC ACATTCATCA TCCAGGCCTA TGACTGTGGT GCTGGGCCCC ACGAGACAGC 420
CTGGAAAAAG TCACACAAGG CCGTGGTCCA TATACAGGTG AAGGATGTCA ACGAGTTTGC 480
TCCCACCTTC AAAGAGCCAG CCTACAAGGC TGTTGTGACG GAGGGCAAGA TCTATGACAG 540
CATTCTGCAG GTGGAGGCCA TTGACGAGGA CTGCTCCCCA CAGTACAGCC AGATCTGCAA 600
CTATGAAATC GTCACCACAG ATGTGCCTTT TGCCATCGAC AGAAATGGCA ACATCAGGAA 660
CACTGAGAAG CTGAGCTATG ACAAACAACA CCAGTATGAG ATCCTGGTGA CCGCCTACGA 720
CTGTGGACAG AAGCCCGCTG CTCAGGACAC CCTGGTGCAG GTGGATGTGA AGCCAGTTTG 780
CAAGCCTGGC TGGCAAGACT GGACCAAGAG GATTGAGTAC CAGCCTGGCT CCGGGAGCAT 840

GCCCCTGTTC CCCAGCATCC ACCTGGAGAC GTGCGATGGA GCCGTGTCTT CCCTCCAGAT 900
CGTCACAGAG CTGCAGACTA ATTACATTGG GAAGGGTTGT GACCGGGAGA CCTACTCTGA 960
GAAATCCCTT CAGAAGTTAT GTGGAGCCTC CTCTGGCATC ATTGACCTCT TGCCATCCCC 1020
TAGCGCTGCC ACCAACTGGA CTGCAGGACT GCTGGTGGAC AGCAGTGAGA TGATCTTCAA 1080
GTTTGACGGC AGGCAGGGTG CCAAAATCCC CGATGGGATT GTGCCCAAGA ACCTGACCGA 1140
TCAGTTCACC ATCACCATGT GGATGAAACA CGGCCCCAGC CCTGGTGTGA GAGCCGAGAA 1200
GGAAACCATC CTCTGCAACT CAGACAAAAC CGAAATGAAC CGGCATCACT ATGCCCTGTA 1260
TGTGCACAAC TGCCGCCTCG TCTTTCTCTT GCGGAAGGAC TTCGACCAGG CTGACACCTT 1320
TCGCCCCGCG GAGTTCCACT GGAAGCTGGA TCAGATTTGT GACAAAGAGT GGCACTACTA 1380
TGTCATCAAT GTGGAGTTTC CTGTGGTAAC CTTATACATG GATGGAGCAA CATATGAACC 1440
ATACCTGGTG ACCAACGACT GGCCCATTCA TCCATCTCAC ATAGCCATGC AACTCACAGT 1500
CGGCGCTTGT TGGCAAGGAG GAGAAGTCAC CAAACCACAG TTTGCTCAGT TCTTTCATGG 1560
AAGCCTGGCC AGTCTCACCA TCCGCCCTGG CAAAATGGAA AGCCAGAAGG TGATCTCCTG 1620
CCTGCAGGCC TGCAAGGAAG GGCTGGACAT TAATTCCTTG GAAAGCCTTG GCCAAGGAAT 1680
AAAGTATCAC TTCAACCCCT CGCAGTCCAT CCTGGTGATG GAAGGTGACG ACATTGGGAA 1740
CATTAACCGT GCTCTCCAGA AAGTCTCCTA CATCAACTCC AGGCAGTTCC CAACGGCGGG 1800
TGTGCGGCGC CTCAAAGTAT CCTCCAAAGT CCAGTGCTTT GGGGAAGACG TATGCATCAG 1860
TATCCCTGAG GTAGATGCCT ATGTGATGGT CCTCCAGGCC ATCGAGCCCC GGATCACCCT 1920
CCGGGGCACA GACCACTTCT GGAGACCTGC TGCCCAGTTT GAAAGTGCCA GGGGAGTGAC 1980
CCTCTTCCCT GATATCAAGA TTGTGAGCAC CTTCGCCAAA ACCGAAGCCC CCGGGGACGT 2040
GAAAACCACA GACCCCAAAT CAGAAGTCTT AGAGGAAATG CTTCATAACT TAGATTTCTG 2100
TGACATTTTG GTGATCGGAG GGGACTTGGA CCCAAGGCAG GAGTGCTTGG AGCTCAACCA 2160
CAGTGAGCTC CACCAACGAC ACCTGGATGC CACTAATTCT ACTGCAGGCT ACTCCATCTA 2220
CGGTGTGGGC TCCATGAGCC GCTATGAGCA GGTGCTACAT CACATCCGCT ACCGCAACTG 2280
GCGTCCGGCT TCCCTTGAGG CCCGGCGTTT CCGGATTAAG TGCTCAGAAC TCAATGGGCG 2340
CTACACTAGC AATGAGTTCA ACTTGGAGGT CAGCATCCTT CATGAAGACC AAGTCTCAGA 2400
TAAGGAGCAT GTCAATCATC TGATTGTGCA GCCTCCCTTC CTCCAGTCTG TCCATCATCC 2460
TGAGTCCCGG AGTAGCATCC AGCACAGTTC AGTGGTCCCA AGCATTGCCA CAGTGGTCAT 2520
CATCATCTCC GTGTGCATGC TTGTGTTTGT CGTGGCCATG GGTGTGTACC GGGTCCGGAT 2580
CGCCCACCAG CACTTCATCC AGGAGACTGA GGCTGCCAAG GAATCTGAGA TGGACTGGGA 2640
CGATTCTGCG CTGACTATCA CAGTCAACCC CATGGAGAAA CATGAAGGAC CAGGGCATGG 2700
GGAAGATGAG ACTGAGGGAG AAGAGGAGGA AGAAGCCGAG GAAGAAATGA GCTCCAGCAG 2760
TGGCTCTGAC GACAGCGAAG AGGAGGAGGA GGAGGAAGGG ATGGGCAGAG GCAGACATGG 2820
GCAGAATGGA GCCAGGCAAG CCCAGCTGGA GTGGGATGAC TCCACCCTCC CCTACTAGTG 2880
CCCAGGGGTC TGCTGCCTGG CCCACATGTC CCTTTTGTAA ACCCTGACCC AGTGTATGCC 2940
CATGTCTATC ATACCTCACC TCTGATGTCT GTGACATGTC TGGGAAGGCC TTCTCCAGCT 3000
TCCTGGAGCC CACCCTTTAA GCCTTGGGCA CTCCCTGTGT TTCATCCATG GGGAAGTTCC 3060
AAGAAGCCCA GCATGGCCAT CAGTGAGGAC TTCAGGGTAG ACTTTGTCCT GTAGCCTCCA 3120
CTTCTGCCCT AAGTTCCCCA GCATCCTGAC TACCTGTCTG CAGAGTTTGC CTTTGTTTTT 3180
TCCTGCAGGG AAGAAGGCCC ACCTTTGTGT CACTCACCTC CCCAGGCTCA GAGTCCCCAA 3240
GGCCCTGGGG TTCCAACTCA CTGTGCGTCT CCTCCACACA GACCAGTAGG TTCTCCTATG 3300
CTGACTCCAG GTTGCTTCAT ACAAGGAGGG TGGTTGAACT TCACACACGT AAGGTCTTAG 3360
TGCTTAACAG TTTAAAGGAA AGTCCTTGTT GAGGCAGAAC TAAGTTTACA GGGAAAGGTA 3420
CACACATTCT CTCTCTCTCT CTCTCTCTGT CTATCTAGTT CCCCAGCTTG GAGAGCCTTT 3480
CCCCTTGCTT CTTTCTGAGG CCATATAAGC TTATAAGAAA AGTCCCAAAC CAAGAATAGG 3540
TCCTTGGCCA CAAGCAGGGT CTGATCCCCC ATCAGAGCTA TCTGAGCCTG CCTGTCTGGG 3600
CACCTGCTGC AACCATGCAG CTACCCTGCC AGGGGCACTC AGCAAACAGA ACCACAGGGC 3660
CCAGGAGGCA TTCCACACAG GCACTGCCCC AGGACAACAC AACAAGGACA GTCACAACAA 3720
GGACAACAAG GACACAACAC AACACACAAC AAGGACAGTC ACAACAAGCC TAGAGCCAGA 3780
AAGCAGATGG AAATGCTAAT GAGGTCAAAC GTAGGCTTCA TGGTGGGTGG AGTGGGGGTG 3840
GCTGGGCTCC CCCAGGACAG AGGGGACCCT GAGGTTGGCA AGGCTCTCAC CACTCAGCCT 3900
TATGGTCCCT TATCTCCTAT CTTCCCTCTT GAGAAAATAC ACGCTTTCTG CATGTATTAG 3960
AAACGCACGA GCTCCACCAA GTCTACAATG AAAGTTTGAA ATTTAACTGC AAGGAATTAG 4020
AAGCATATTT GCAATCATTG CAGCTTCTTC TTTCTTCTGC TCATAAAAGG AGGAACACTT 4080
TAGATAGAGG GCAAATATAT CTGAAAACCT AATTTCTTTC TTTTTTTGAT AAGGAAATCT 4140
TTTCCATCTC CATCCTAACA TGCACAACCT GTGAAGAGAA TTGTTTCTAT AGTAACTGGT 4200
CTGTGATCTT TTGTGGCCAA GAGAATAGCA GGCAAGAATT AGGGCCTTGA CAGAATTTCC 4260
ACGAAGCTCT GAGAACATGT TTGTTTCGAA TGTCTGATTC CTCTTTGTCA TCAATGTGTA 4320
TGCTCTGTCC CCATCCTTCA CTCCTCCTCA AGCTCACACC AATTGGTTTG GCACAGGCAC 4380
AGAGCTGGTC CCTAGTTAAG TGGCATTTAT GTTAAAAAAA A

BFA1 Protein sequence
Gene name: calsyntenin-2; Unigene number: Hs.7413; Probeset Accession #: R46025; Protein Accession #: NP_071414; Predicted Signal sequence: 1-20; Predicted TM domains: 832-848; PFAM domains: cadherin_domains: 48-151, 165-254; Summary: A type I membrane protein; a member of the calsyntenin family; is related to the FAT tumor suppressor; is likely an adhesion molecule important in mammalian developmental processes and cell communication.

MLPGRLCWVP LLLALGVGSG SGGGGDSRQR RLLAAKVNKH KPWIETSYHG VITENNDTVI 60
LDPPLVALDK DAPVPFAGEI CAFKIHGQEL PFEAVVLNKT SGEGRLRAKS PIDCELQKEY 120
TFIIQAYDCG AGPHETAWKK SHKAVVHIQV KDVNEFAPTF KEPAYKAVVT EGKIYDSILQ 180
VEAIDEDCSP QYSQICNYEI VTTDVPFAID RNGNIRNTEK LSYDKQHQYE ILVTAYDCGQ 240

KPAAQDTLVQ VDVKPVCKPG WQDWTKRIEY QPGSGSMPLF PSIHLETCDG AVSSLQIVTE 300
LQTNYIGKGC DRETYSEKSL QKLCGASSGI IDLLPSPSAA TNWTAGLLVD SSEMIFKFDG 360
RQGAKIPDGI VPKNLTDQFT ITMWMKHGPS PGVRAEKETI LCNSDKTEMN RHHYALYVHN 420
CRLVFLLRKD FDQADTFRPA EFHWKLDQIC DKEWHYYVIN VEFPVVTLYM DGATYEPYLV 480
TNDWPIHPSH IAMQLTVGAC WQGGEVTKPQ FAQFFHGSLA SLTIRPGKME SQKVISCLQA 540
CKEGLDINSL ESLGQGIKYH FNPSQSILVM EGDDIGNINR ALQKVSYINS RQFPTAGVRR 600
LKVSSKVQCF GEDVCISIPE VDAYVMVLQA IEPRITLRGT DHFWRPAAQF ESARGVTLFP 660
DIKIVSTFAK TEAPGDVKTT DPKSEVLEEM LHNLDFCDIL VIGGDLDPRQ ECLELNHSEL 720
HQRHLDATNS TAGYSIYGVG SMSRYEQVLH HIRYRNWRPA SLEARRFRIK CSELNGRYTS 780
NEFNLEVSIL HEDQVSDKEH VNHLIVQPPF LQSVHHPESR SSIQHSSVVP SIATVVIIIS 840
VCMLVPVVAM GVYRVRIAHQ HFIQETEAAK ESEMDWDDSA LTITVNPMEK HEGPGHGEDE 900
TEGEEEEEAE EEMSSSSGSD DSEEEEEEEG MGRGRHGQNG ARQAQLEWDD STLPY

BFG7 DNA sequence
Gene name: EST; Unigene number: Hs.91668; Probeset Accession #: Z40805; Nucleic Acid Accession #: n/a; Coding sequence: <1-906 (stop codon underlined)

CGGGTCGACC CACGCGTCCG GGGAGAAAGG ATGGCCGGCC TGGCGGCGCG GTTGGTCCTG 60
CTAGCTGGGG CAGCGGCGCT GGCGAGCGGC TCCCAGGGCG ACCGTGAGCC GGTGTACCGC 120
GACTGCGTAC TGCAGTGCGA AGAGCAGAAC TGCTCTGGGG GCGCTCTGAA TCACTTCCGC 180
TCCCGCCAGC CAATCTACAT GAGTCTAGCA GGCTGGACCT GTCGGGACGA CTGTAAGTAT 240
GAGTGTATGT GGGTCACCGT TGGGCTCTAC CTCCAGGAAG GTCACAAAGT GCCTCAGTTC 300
CATGGCAAGT GGCCCTTCTC CCGGTTCCTG TTCTTTCAAG AGCCGGCATC GGCCGTGGCC 360
TCGTTTCTCA ATGGCCTGGC CAGCCTGGTG ATGCTCTGCC GCTACCGCAC CTTCGTGCCA 420
GCCTCCTCCC CCATGTACCA CACCTGTGTG GCCTTCGCCT GGGTGTCCCT CAATGCATGG 480
TTCTGGTCCA CAGTYTTCCA CACCAGGGAC ACTGACCTCA CAGAGAAAAT GGACTACTTC 540
TGTGCCTCCA CTGTCATCCT ACACTCAATC TACCTGTGCT GCGTCAGCCT CATCCGCTTC 600
GACTATGGCT ACAACCTGGT GGCCAACGTG GCTATTGGCC TGGTCAACGT GGTGTGGTGG 660
CTGGCCTGGT GCCTGTGGAA CCAGCGGCGG CTGCCTCACG TGCGCAAGTG CGTGGTGGTG 720
GTCTTGCTGC TGCAGGGGCT GTCCCTGCTC GAGCTGCTTG ACTTCCCACC GCTCTTCTGG 780
GTCCTGGATG CCCATGCCAT CTGGCACATC AGCACCATCC CTGTCCACGT CCTCTTTTTC 840
AGCTTTCTGG AAGATGACAG CCTGTACCTG CTGAAGGAAT CAGAGGACAA GTTCAAGCTG 900
GACTGAAGAC CTTGGAGCGA GTCTGCCCCA GTGGGGATCC TGCCCCCGCC CTGCTGGCCT 960
CCCTTCTCCC CTCAACCCTT GAGATGATTT TCTCTTTTCA ACTTCTTGAA CTTGGACATG 1020
AAGGATGTGG GCCCAGAATC ATGTGGCCAG CCCACCCCCT GTTGGCCCTC ACCAGCCTTG 1080
GAGTCTGTTC TAGGGAAGGC CTCCCAGCAT CTGGGACTCG AGAGTGGGCA GCCCCTCTAC 1140
CTCCTGGAGC TGAACTGGGG TGGAACTGAG TGTGCTCTTA GCTCTACCGG GAGGACAGCT 1200
GCCTGTTTCC TCCCCATCAG CCTCCTCCCC ACATCCCCAG CTGCCTGGCT GGGTCCTGAA 1260
GCCCTCTGTC TACCTGGGAG ACCAGGGACC ACAGGCCTTA GGGATACAGG GGGTCCCCTT 1320
CTGTTACCAC CCCCCACCCT CCTCCAGGAC ACCACTAGGT GGTGCTGGAT GCTTGTTCTT 1380
TGGCCAGCCA AGGTTCACGG CGATTCTCCC CATGGGATCT TGAGGGACCA AGCTGCTGGG 1440
ATTGGGAAGG AGTTTCACCC TGACCRTTGC CCTAGCCAGG TTCCCAGGAG GCCTCACCAT 1500
ACTCCCTTTC AGGGCCAGGG CTCCAGCAAG CCCAGGGCAA GGATCCTGTG CTGCTGTCTG 1560
GTTGAGAGCC TGCCACCGTG TGTCGGGAGT GTGGGCCAGG CTGAGTGCAT AGGTGACAGG 1620
GCCGTGAGCA TGGGCCTGGG TGTGTGTGAG CTCAGGCACT AGGTGCGCAG TGTGGAGACG 1680
GGTGTTGTCG GGGAAGAGGT GTGGCTTCAA AGTGTTGTGT GTGCAGGGGG TKGGTGTGTT 1740
AAGCGTGGGT TAGGGGAACG TGTGTGCGCG TGCTGGTGGG CATGTGAGAT GAGTGACTGC 1800
CGGTGAATGT GTCCACAGTT GAGAGGTTGG AGCAGGATGA GGGAATCCTG TCACCATCAA 1860
TAATCACTTG TGGAGCGCCA CTTGGCCCAA GACGCCACCT GGGCGGACAG CAGGAGCTCT 1920
CCATGGCCAG GCTGCCTGTG TGCATGTTCC CTGTCTGGTG CCCCTTTGCC CGCCTCCTGC 1980
AAACCTCACA GGGTCCCCAC ACAACAGTGC CCTCCAGAAG CAGCCCCTCG GAGGCAGAGG 2040
AAGGAAAATG GGGATGGCTG GGGCTCTCTC CATCCTCCTT TTCTCCTTGC CTTCGCATGG 2100
CTGGCCTTCC CCTCCAAAAC CTCCATTCCC CTGCTGCCAG CCCCTTTGCC ATAGCCTGAT 2160
TTTGGGGAGG AGGAAGGGGC GATTTGAGGG AGAAGGGGAG AAAGCTTATG GCTGGGTCTG 2220
GTTTCTTCCC TTCCCAGAGG GTCTTACTGT TCCAGGGTGG CCCCAGGGCA GGCAGGGGCC 2280
ACACTATGCC TGCGCCCTGG TAAAGGTGAC CCCTGCCATT TACCAGCAGC CCTGGCATGT 2340
TCCTGCCCCA CAGGAATAGA ATGGAGGGAG CTCCAGAAAC TTTCCATCCC AAAGGCAGTC 2400
TCCGTGGTTG AAGCAGACTG GATTTTTGCT CTGCCCCTGA CCCCTTGTCC CTCTTTGAGG 2460
GAGGGGAGCT ATGCTAGGAC TCCAACCTCA GGGACTCGGG TGGCCTGCGC TAGCTTCTTT 2520
TGATACTGAA AACTTTTAAG GTGGGAGGGT GGCAAGGGAT GTGCTTAATA AATCAATTCC 2580
AAGCCTCAAA AAAAAAAAAA AAAAAAAAAA AAAAAA

BFG7 Protein sequence
Gene name: EST; Unigene number: Hs.91668; Probeset Accession #: Z40805; Protein Accession #: n/a; Signal sequence: none; Predicted TM domains: 117-133, 179-195, 211-227, 235-251, 266-282, 296-312; PFAM domains: none; Summary: A type III membrane protein of unknown function; is adjacent to HER2 on the genome, and its overexpression in breast cancer is highly correlated with HER2 expression; may be used to predict HER2 overexpression and amplification.

RVDPRVRGER MAGLAARLVL LAGAAALASG SQGDREPVYR DCVLQCEEQN CSGGALNHFR 60
SRQPIYMSLA GWTCRDDCKY ECMWVTVGLY LQEGHKVPQF HGKWPFSRFL FFQEPASAVA 120
SFLNGLASLV MLCRYRTFVP ASSPMYHTCV AFAWVSLNAW FWSTVFHTRD TDLTEKMDYF 180

```
CASTVILHSI YLCCVRTVGL QHPAVVSAFR ALLLLMLTVH VSYLSLIRFD YGYNLVANVA   240
IGLVNVVWWL AWCLWNQRRL PHVRKCVVVV LLLQGLSLLE LLDFPPLFWV LDAHAIWHIS   300
TIPVHVLFFS FLEDDSLYLL KESEDKFKLD

BCN4 DNA sequence
Gene name:  ESTs; Unigene number:  Hs.283713; Probeset Accession #:  F13673; Nucleic Acid
Accession #:  n/a; Coding sequence:  143-874 (start and stop codons underlined)

GGGAGGGAGA GAGGCGCGCG GGTGAAAGGC GCATTGATGC AGCCTGCGGC GGCCTCGGAG    60
CGCGGCGGAG CCAGACGCTG ACCACGTTCC TCTCCTCGGT CTCCTCCGCC TCCAGCTCCG   120
CGCTGCCCGG CAGCCGGGAG CCATGCGACC CCAGGGCCCC GCCGCCTCCC CGCAGCGGCT   180
CCGCGGCCTC CTGCTGCTCC TGCTGCTGCA GCTGCCCGCG CCGTCGAGCG CCTCTGAGAT   240
CCCCAAGGGG AAGCAAAAGG CGCAGCTCCG GCAGAGGGAG GTGGTGGACC TGTATAATGG   300
AATGTGCTTA CAAGGGCCAG CAGGAGTGCC TGGTCGAGAC GGGAGCCCTG GGGCCAATGG   360
CATTCCGGGT ACACCTGGGA TCCCAGGTCG GGATGGATTC AAAGGAGAAA AGGGGGAATG   420
TCTGAGGGAA AGCTTTGAGG AGTCCTGGAC ACCCAACTAC AAGCAGTGTT CATGGAGTTC   480
ATTGAATTAT GGCATAGATC TTGGGAAAAT TGCGGAGTGT ACATTTACAA AGATGCGTTC   540
AAATAGTGCT CTAAGAGTTT TGTTCAGTGG CTCACTTCGG CTAAAATGCA GAAATGCATG   600
CTGTCAGCGT TGGTATTTCA CATTCAATGG AGCTGAATGT TCAGGACCTC TTCCCATTGA   660
AGCTATAATT TATTTGGACC AAGGAAGCCC TGAAATGAAT TCAACAATTA ATATTCATCG   720
CACTTCTTCT GTGGAAGGAC TTTGTGAAGG AATTGGTGCT GGATTAGTGG ATGTTGCTAT   780
CTGGGTTGGC ACTTGTTCAG ATTACCCAAA AGGAGATGCT TCTACTGGAT GGAATTCAGT   840
TTCTCGCATC ATTATTGAAG AACTACCAAA ATAAATGCTT TAATTTTCAT TTGCTACCTC   900
TTTTTTTATT ATGCCTTGGA ATGGTTCACT TAAATGACAT TTTAAATAAG TTTATGTATA   960
CATCTGAATG AAAAGCAAAG CTAAATATGT TTACAGACCA AAGTGTGATT TCACACTGTT  1020
TTTAAATCTA GCATTATTCA TTTTGCTTCA ATCAAAAGTG GTTTCAATAT TTTTTTTAGT  1080
TGGTTAGAAT ACTTTCTTCA TAGTCACATT CTCTCAACCT ATAATTTGGA ATATTGTTGT  1140
GGTCTTTTGT TTTTTCTCTT AGTATAGCAT TTTTAAAAAA ATATAAAAGC TACCAATCTT  1200
TGTACAATTT GTAAATGTTA AGAATTTTTT TTATATCTGT TAAATAAAAA TTATTTCCAA  1260
CAACCTTAAA AAAAAAAAAA AAAA

BCN4 Protein sequence
Gene name:  ESTs; Unigene number:  Hs.283713; Probeset Accession #:  F13673; Protein Accession
#:  n/a; Predicted Signal sequence:  1-30; TM domains:  none; PFAM domains:  none; Summary:  a
secreted protein; has a mouse orthologue (see sequence below).

MRPQGPAASP QRLRGLLLLL LLQLPAPSSA SEIPKGKQKA QLRQREVVDL YNGMCLQGPA    60
GVPGRDGSPG ANGIPGTPGI PGRDGFKGEK GECLRESFEE SWTPNYKQCS WSSLNYGIDL   120
GKIAECTFTK MRSNSALRVL FSGSLRLKCR NACCQRWYFT FNGAECSGPL PIEAIIYLDQ   180
GSPEMNSTIN IHRTSSVEGL CEGIGAGLVD VAIWVGTCSD YPKGDASTGW NSVSRIIIEE   240
LPK

Mouse BCN4 Protein sequence
Gene name:  ESTs; Unigene number:  Mm.41556

XXXXAAPPQL LLGLFLVLLL LLQLSAPSSA SENPKVKQKA LIRQREVVDL YNGMCLQGPA    60
GVPGRDGSPG ANGIPGTPGI PCQDGFKGEK GECLRESFEE SWTPNYKQCS WSSLNYGIDL   120
GKIAECTFTK MRSNSALRVL FSGSLRLKCR NACCQRWYFT FNGAECSGPP PIEAIXXXXX   180
XXXXXXXXXX XXXXXXXXXX XXXXXXXXXX XXXXXXXXSD YPKGDAYTGW DSVSRIIIEE   240
LPK
```

It is understood that the examples described above in no way serve to limit the true scope of this invention, but rather are presented for illustrative purposes. All publications, sequences of accession numbers, and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference.